Guided Zeroth-Order Methods for Stochastic Non-convex Problems with Decision-Dependent Distributions

Yuya Hikima [1] **Hiroshi Sawada** [1] **Akinori Fujino** [1]

Abstract

In this study, we tackle an optimization problem with a known function and an unknown decision-dependent distribution, which arises in a variety of applications and is often referred to as a performative prediction problem. To solve the problem, several zeroth-order methods have been developed because the gradient of the objective function cannot be obtained explicitly due to the unknown distribution. Although these methods have theoretical convergence, they cannot utilize the information on the known function, which limits their efficiency in reducing the objective value. To overcome this issue, we propose new zeroth-order methods that generate effective update directions by utilizing information on the known function. As theoretical results, we show the convergence of our methods to stationary points and provide the worst-case sample complexity analysis. Our simulation experiments on multiple applications show that our methods output solutions with lower objective values than the existing zeroth-order methods do.

1. Introduction

In this study, we consider the following problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad F(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x})}[f(\boldsymbol{x},\boldsymbol{\xi})], \tag{1}$$

where $F:\mathbb{R}^d\to\mathbb{R}$ is generally non-convex, f is a given differentiable function, and $D(\boldsymbol{x})$ is an unknown distribution of random variables $\boldsymbol{\xi}$. This formulation is known as the *performance prediction problem* (Perdomo et al., 2020). The main feature of this problem is that the probability distribution $D(\boldsymbol{x})$ depends on decision vector $\boldsymbol{x}$ and is unknown. This problem appears in a wide range of applications. For example, in pricing applications, a seller aims to optimize the prices $\boldsymbol{x}$ of products and services for their revenue, considering the *price-dependent* distribution $D(\boldsymbol{x})$ of stochastic demand (Ray et al., 2022; Hikima & Takeda, 2025b). In strategic classification in financial practices, a lender aims to train the parameter $\boldsymbol{x}$ of a classifier for finding good customers, considering the *parameter-dependent* data distribution $D(\boldsymbol{x})$, which is caused by customers who may react to the deployed classifier (Levanon & Rosenfeld, 2021; Liu et al., 2024a). Various studies (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Chen et al., 2024) have addressed problem (1). because of its importance in practical applications.

Since the gradient of the objective function is inaccessible as a result of the distribution $D(\boldsymbol{x})$ being unknown, various zeroth-order methods have been proposed, which update the decision vector by some estimated gradient on the basis of function evaluation(s). For example, (Chen et al., 2024; Hikima & Takeda, 2025b) proposed zeroth-order methods based on the following gradient estimator:

$$\boldsymbol{g} := \frac{1}{2\mu}\left(f(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{\xi}^1) - f(\boldsymbol{x}-\mu\boldsymbol{u},\boldsymbol{\xi}^2)\right)\boldsymbol{u}, \tag{2}$$

where $\mu\in\mathbb{R}_{\geq 0}$ is a constant, $\boldsymbol{u}\in\mathbb{R}^d$ is sampled from some distribution (e.g., the standard Gaussian distribution), $\boldsymbol{\xi}^1\sim D(\boldsymbol{x}+\mu\boldsymbol{u})$, and $\boldsymbol{\xi}^2\sim D(\boldsymbol{x}-\mu\boldsymbol{u})$. Not limited to the above, various studies (Ray et al., 2022; Liu et al., 2024a) proposed zeroth-order methods using different gradient estimators.

Although these existing zeroth-order methods are theoretically valid for solving problem (1), they treat the function f as a black box and do not leverage its gradient information, which limits optimization efficiency. Specifically, an unbiased stochastic gradient of the objective function F in (1) can be written as follows (Liu et al., 2024b; Hikima & Takeda, 2025a):

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{\xi}) + f(\boldsymbol{x},\boldsymbol{\xi})\nabla_{\boldsymbol{x}}\log\Pr(\boldsymbol{\xi}\mid\boldsymbol{x}),\ \boldsymbol{\xi}\sim D(\boldsymbol{x}). \tag{3}$$

Here, although we cannot obtain the second term since $\Pr(\boldsymbol{\xi}\mid\boldsymbol{x})$ is unknown, we can calculate the first term, that is, $\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{\xi})$, since f is known. Thus, we can access partial information on the gradient of objective function F.

[1]Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Correspondence to: Yuya Hikima <yuya.hikima@ntt.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

The fact that existing methods do not fully utilize the information leads to slow convergence and increases the number of samples $\boldsymbol{\xi} \sim D(\boldsymbol{x})$. Such an increase is undesirable in practice because a decision ($\boldsymbol{x}$) must be deployed in real-world to obtain a sample $\boldsymbol{\xi} \sim D(\boldsymbol{x})$. For example, a decision maker must sell some products at prices $\boldsymbol{x}$ to obtain sample demand $\boldsymbol{\xi} \sim D(\boldsymbol{x})$.

In this work, we propose guided zeroth-order methods that integrate partial gradient information derived from the known structure of f. Inspired by (Maheswaranathan et al., 2019), our methods use the gradient estimator (2) with $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\mathcal{N}(0, \boldsymbol{\Sigma})$ is a normal distribution with a variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. The key to our methods is to incorporate partial gradient information in $\boldsymbol{\Sigma}$ in order to generate an updating direction $\boldsymbol{u}$ that tends to decrease the objective value. We design two interchangeable methods for obtaining partial gradient information. One method computes, as the information, the first term of (3) from new sample(s) $\boldsymbol{\xi} \sim D(\boldsymbol{x}_k)$ at the current iteration k. The other method approximates it from historical sample(s) $\boldsymbol{\xi}_i \sim D(\boldsymbol{x}_i)$ in past iteration(s) i.

Our theoretical analysis shows that our methods converge to stationary points and provides the worst-case *sample complexity*, which represents the number of samples $\boldsymbol{\xi} \sim D(\boldsymbol{x})$ needed to obtain a stationary point $\hat{\boldsymbol{x}}$ such that $\mathbb{E}[\|\nabla F(\hat{\boldsymbol{x}})\|^2] \leq \epsilon^2$. In particular, we show that the worst-case sample complexity of our methods is $O(\sigma^3 d^4 \epsilon^{-6})$, where $\mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x})}[(F(\boldsymbol{x}) - f(\boldsymbol{x}, \boldsymbol{\xi}))^2] \leq \sigma^2$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Since the sample complexities of the existing methods are $O(\sigma^3 d^{\frac{9}{2}} \epsilon^{-6})$ (Hikima & Takeda, 2025b) and $O(G^6 d^2 \epsilon^{-6})$ (Liu et al., 2024a), where $G = \sup_{\boldsymbol{x},\boldsymbol{\xi}} |f(\boldsymbol{x}, \boldsymbol{\xi})|$, our methods outperform them when $G = \omega(d^{\frac{1}{3}})$ or G is unbounded. Note that even if f is a simple loss function such as squared error, G is unbounded. This improved order is achieved by employing a tighter convergence analysis compared to (Hikima & Takeda, 2025b), which involves tuning algorithmic parameters such as mini-batch size and step size. Furthermore, the proposed zeroth-order methods incorporate partial gradient information without making any assumptions about its correlation with the true gradient, and crucially, without deteriorating the convergence rate.

We conducted simulation experiments on multiple products pricing and strategic classification applications. The results show that our methods output solutions with lower objective values than the existing zeroth-order methods do.

Notation. Bold lowercase symbols (e.g., $\boldsymbol{x}, \boldsymbol{y}$) denote vectors, and $\|\boldsymbol{x}\|$ denotes the Euclidean norm of a vector $\boldsymbol{x}$. Bold uppercase symbols (e.g., $\boldsymbol{\Sigma}$) denote matrices, and $\|\boldsymbol{\Sigma}\|$ denotes the spectral norm of a matrix $\boldsymbol{\Sigma}$. The inner product of the vectors $\boldsymbol{x}, \boldsymbol{y}$ is denoted by $\boldsymbol{x}^\top \boldsymbol{y}$. Let $\mathbb{R}_{>0}$ ($\mathbb{R}_{\geq 0}$) be the set of positive (non-negative) real numbers. Let $\mathbb{N}$ be the set of natural numbers. The gradient for a real-valued function $f(\boldsymbol{x})$ *w.r.t.* $\boldsymbol{x}$ is denoted by $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$. We let $\boldsymbol{I}_d$ be a d-dimensional identity matrix and let $\mathcal{N}(0, \boldsymbol{\Sigma})$ be a Gaussian distribution with a co-variance matrix $\boldsymbol{\Sigma}$. We denote the set of $d \times d$ positive definite matrices by $\mathbb{S}^d_{++}$. Let $[N]$ be the set $\{1, 2, \ldots, N\}$.

2. Related Work

2.1. Zeroth-Order Optimization Methods

Zeroth-order optimization methods have been proposed for solving optimization problems where gradient and Hessian information are either unavailable or computationally expensive to calculate (Powell, 2006; Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017; Ragonneau & Zhang, 2024). Recently, various studies have proposed zeroth-order methods for solving problems with decision-dependent distributions (Ray et al., 2022; Liu et al., 2024a; Chen et al., 2024; Hikima & Takeda, 2025b). (Chen et al., 2024) derived the conditions under which (1) can be reduced to a convex optimization problem and solved the reduced convex problem by using a zeroth-order method. (Hikima & Takeda, 2025b) proposed zeroth-order methods with one-point and two-point gradient estimators and derived the iteration and sample complexities for reaching stationary points. (Ray et al., 2022; Liu et al., 2024a) proposed zeroth-order methods for (1) with time-varying decision-dependent distributions. Although these studies effectively develop zeroth-order methods within their settings, they fail to leverage the known structure of the function f. In this study, we propose new zeroth-order methods by using the information on f to generate updating directions.

2.2. Other Methods of Solving Stochastic Problems with Decision-Dependent Distributions

Here, we list methods other than zeroth-order ones for (1).

Retraining methods (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Roy et al., 2022; Mofakhami et al., 2023; Li & Wai, 2024; Khorsandi et al., 2024). Retraining methods update the decision vector without accounting for the distribution shift at each iteration. Specifically, *repeated gradient descent* (Perdomo et al., 2020) updates the decision vector such that $\boldsymbol{x}_{k+1} := \mathrm{proj}_{\mathcal{C}}(\boldsymbol{x}_k - \eta_k \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x}_k)}[\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{\xi})])$, where $\mathcal{C}$ is the feasible region and $\mathrm{proj}_{\mathcal{C}}$ is the Euclidean projection operator onto $\mathcal{C}$. This method makes the decision vector to converge to a *performatively stable point* $\boldsymbol{x}_{\mathrm{PS}} = \arg\min_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x}_{\mathrm{PS}})}[f(\boldsymbol{x}, \boldsymbol{\xi})]$. Subsequent studies have explored several directions, including variants of the above approach (Mendler-Dünner et al., 2020), a setting where the data distribution follows a Markov chain with a state-dependent transition kernel (Roy et al., 2022), relaxed assumptions on the objective function

(Mofakhami et al., 2023; Li & Wai, 2024), and improved convergence analyses (Khorsandi et al., 2024). However, these methods focus on obtaining performatively stable solutions and do not aim to (nor is capable of) finding stationary points, which may result in worse function values.

Stochastic first-order methods (Liu et al., 2024b; Hikima & Takeda, 2025a). These methods assume that $\Pr(\boldsymbol{\xi} \mid \boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} \Pr(\boldsymbol{\xi} \mid \boldsymbol{x})$ are known and update the decision vector by using the unbiased stochastic gradient (3) of the objective function. However, these methods cannot be used for solving (1) with an unknown distribution $D(\boldsymbol{x})$.

Performative gradient descent (Izzo et al., 2021). This method updates the iterates using an approximate gradient obtained by estimating $\nabla_{\boldsymbol{x}} \log \Pr(\boldsymbol{\xi} \mid \boldsymbol{x})$ in the second term of (3). While this approach can find an optimal point rather than a performative stable point, it relies on several restrictive assumptions. For example, they assume that the distribution can be written as $\Pr(\boldsymbol{\xi} \mid \boldsymbol{x}) = p(q(\boldsymbol{x}), \boldsymbol{\xi})$, where the function p is known and the intermediate parameter q can be approximately obtained through samples. They also assume that the objective function F is convex. In contrast, our study suppose looser assumptions: the distribution $D(\boldsymbol{x})$ is completely unknown and the objective function is generally nonconvex.

Distribution approximating approaches (Miller et al., 2021; Lin & Zrnic, 2024). These approaches estimate models of the distribution map $D(\cdot)$ and optimize problem (1) by using the estimated distribution. While these approaches perform well when the distribution is approximated accurately, they assume specific distributions or functions. For example, (Miller et al., 2021) assume that the distribution map is included in location-scale families, i.e., $D(\boldsymbol{x})$ needs to satisfy $\boldsymbol{\xi} \sim D(\boldsymbol{x}) \Leftrightarrow \boldsymbol{\xi} := \boldsymbol{\xi}_0 + \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{\xi}_0$ is a random variable independent of $\boldsymbol{x}$ and matrix $\boldsymbol{A}$ is constant. Moreover, they assume that f is strongly convex. In this paper, we tackle the design of (more) generic methods for solving (1) without such specific assumptions.

Search optimization (Bergstra & Bengio, 2012; Frazier, 2018; Xue & Shen, 2020). Search optimization techniques, such as Bayesian optimization (Frazier, 2018), random search (Bergstra & Bengio, 2012), and sparrow search algorithm (Xue & Shen, 2020), aim to find a global solution by iteratively evaluating objective values across the search space. When these methods are applied to our problem, they require a large number of samples $\boldsymbol{\xi} \sim D(\boldsymbol{x})$ to obtain the objective values. Having to take such a large number of samples is undesirable from the perspective of the sample complexity, as noted in the introduction.

3. Proposed Method

Here, we describe our new gradient estimator in Section 3.1 and the algorithm using the gradient estimator in Section 3.2. Moreover, we describe an advanced algorithm that utilizes historical samples from past iterations in Section 3.3.

3.1. Our Gradient Estimator

We propose the following gradient estimator:

$$\boldsymbol{g}(\boldsymbol{x}, \mu, \boldsymbol{u}, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2) := \frac{f(\boldsymbol{x} + \mu\boldsymbol{u}, \boldsymbol{\xi}^1) - f(\boldsymbol{x} - \mu\boldsymbol{u}, \boldsymbol{\xi}^2)}{2\mu}\boldsymbol{u}, \tag{4}$$

where $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, $\boldsymbol{\xi}^1 \sim D(\boldsymbol{x} + \mu\boldsymbol{u})$, $\boldsymbol{\xi}^2 \sim D(\boldsymbol{x} - \mu\boldsymbol{u})$, and $\mu \in \mathbb{R}_{>0}$. Here, we utilize partial information on the gradient to set $\boldsymbol{\Sigma}$, inspired by (Maheswaranathan et al., 2019). Specifically, we regard the following vector as partial gradient information:

$$\boldsymbol{s} := \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\xi}^i), \tag{5}$$

where $\boldsymbol{\xi}^i \sim D(\boldsymbol{x})$ for $i \in \{1, 2, \ldots, n\}$. This is because $\boldsymbol{s}$ approximates the expectation of the first term of the unbiased stochastic gradient (3) for the objective function F. Let $\boldsymbol{h}$ be the normalized vector of $\boldsymbol{s}$, that is, $\boldsymbol{h} = \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|}$ if $\boldsymbol{s} \neq \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{0}$ otherwise. Accordingly, we define $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\Sigma} := \frac{\alpha}{d}\boldsymbol{I}_d + (1 - \alpha)\boldsymbol{h}\boldsymbol{h}^\top, \tag{6}$$

where $\alpha \in (0, 1]$. Here, α is a balance parameter that adjusts the weight of the partial gradient information.

Next, we present the property of our gradient estimator. First, let us define an important function, which is closely related to our gradient estimator.

Definition 3.1. We call the following function the augmented Gaussian smoothed function of F.

$$F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[F(\boldsymbol{x} + \mu\boldsymbol{u})].$$

The function $F_{\mu,\boldsymbol{\Sigma}}$ serves as a smooth approximation of F. This is an extension of the *Gaussian smoothed function* (Nesterov & Spokoiny, 2017; Iwakiri et al., 2022): if $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, function $F_{\mu,\boldsymbol{\Sigma}}$ is consistent with it. Throughout the paper, we assume that $F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})$ is well-defined, that is, $\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[|F(\boldsymbol{x} + \mu\boldsymbol{u})|] < \infty$ for any given $\mu \in \mathbb{R}_{>0}$, positive-definite matrix $\boldsymbol{\Sigma} \in \mathbb{S}^d_{++}$, and $\boldsymbol{x} \in \mathbb{R}^d$.

Next, we show that our gradient estimator is correlated with the gradient of $F_{\mu,\boldsymbol{\Sigma}}$ by the following lemma.

Lemma 3.2. *Suppose that F is Lipschitz continuous. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$, we have*

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\mathbb{E}_{\boldsymbol{\xi}^1\sim D(\boldsymbol{x}+\mu\boldsymbol{u}),\boldsymbol{\xi}^2\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\left[\boldsymbol{g}(\boldsymbol{x}, \mu, \boldsymbol{u}, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2)\right]\right] = \boldsymbol{\Sigma}\nabla_{\boldsymbol{x}} F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}).$$

Algorithm 1 Guided zeroth-order method with new samples

input $\boldsymbol{x}_0 \in \mathbb{R}^d$, $T \in \mathbb{N}$, $\boldsymbol{\Sigma}_0 \in \mathbb{S}^d_{++}$, $\mu_0 \in \mathbb{R}_{>0}$, $\mu_{\min} \in \mathbb{R}_{>0}$, $\eta \in (0,1]$, $\alpha_0 \in (0,1]$, $\gamma \in [0,1)$, $\beta \in \mathbb{R}_{>0}$, $\{m_k\}_{k=0}^T \in \mathbb{N}$, $\{n_k\}_{k=0}^T \in \mathbb{N}$, and distribution D_R for $[T]$.

1: Sample $R \sim D_R(T)$
2: **for** $k = 0, 1, \ldots, R-1$ **do**
3: Sample $\boldsymbol{u}_k$ from $\mathcal{N}(0, \boldsymbol{\Sigma}_k)$, $\{\boldsymbol{\xi}_k^{1,i}\}_{i=1}^{m_k}$ from $D(\boldsymbol{x}_k + \mu_k \boldsymbol{u}_k)$, and $\{\boldsymbol{\xi}_k^{2,i}\}_{i=1}^{m_k}$ from $D(\boldsymbol{x}_k - \mu_k \boldsymbol{u}_k)$.
4: $\boldsymbol{g}_k \leftarrow \frac{1}{m_k} \sum_{i=1}^{m_k} \frac{f(\boldsymbol{x}_k + \mu_k \boldsymbol{u}_k, \boldsymbol{\xi}_k^{1,i}) - f(\boldsymbol{x}_k - \mu_k \boldsymbol{u}_k, \boldsymbol{\xi}_k^{2,i})}{2\mu_k} \boldsymbol{u}_k$.
5: $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \beta \boldsymbol{g}_k$
6: Sample $\{\boldsymbol{\xi}_k^{3,i}\}_{i=1}^{n_k}$ from $D(\boldsymbol{x}_{k+1})$
7: $\boldsymbol{s}_k \leftarrow \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1}, \boldsymbol{\xi}_k^{3,i})$
8: **if** $\boldsymbol{s}_k \neq 0$ **then**
9: $\boldsymbol{\Sigma}_{k+1} \leftarrow \frac{\alpha_k}{d} \boldsymbol{I}_d + (1-\alpha_k) \boldsymbol{h}_k \boldsymbol{h}_k^\top$, where $\boldsymbol{h}_k := \frac{\boldsymbol{s}_k}{\|\boldsymbol{s}_k\|}$
10: **else**
11: $\boldsymbol{\Sigma}_{k+1} \leftarrow \frac{\alpha_k}{d} \boldsymbol{I}_d$
12: **end if**
13: $\mu_{k+1} \leftarrow \max(\eta \mu_k, \mu_{\min})$
14: $\alpha_{k+1} \leftarrow 1 - \gamma(1-\alpha_k)$
15: **end for**
return $\boldsymbol{x}_R$

Note that the proofs of all our lemmas and theorems can be found in Appendix D. Moreover, the Lipschitz condition for F in Lemma 3.2 holds under later Assumptions 4.2 and 4.3 (See Lemma D.2).

From Lemma 3.2, when $\boldsymbol{\Sigma} \neq \boldsymbol{I}_d$, the gradient estimator in (4) is a biased gradient estimator for the augmented Gaussian smoothed function $F_{\mu,\boldsymbol{\Sigma}}$. When $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, it is an unbiased gradient estimator for $F_{\mu,\boldsymbol{\Sigma}}$.

Effect of balance parameter α. From Lemma 3.2 and (6), parameter α adjusts the trade-off between the biasedness of our gradient estimator and the weight of the partial gradient information: if $\alpha = 1$, then $\boldsymbol{\Sigma} = \frac{1}{d}\boldsymbol{I}_d$ and $\boldsymbol{g}$ is an (scaled) unbiased gradient estimator for $F_{\mu,\boldsymbol{\Sigma}}$; if α is close to 0, the distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$ emphasizes the partial gradient information. Our methods (proposed later) initially set α to a small value to promote a faster decrease in the objective value, while in the latter iterations, they make α closer to 1 gradually to obtain a gradient estimator with a small bias.

The way of sampling $\boldsymbol{u}$. Samples of $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ can be generated efficiently as

$$\boldsymbol{u} = \sqrt{\frac{\alpha}{d}} \boldsymbol{w} + \sqrt{1-\alpha} \boldsymbol{h} v, \tag{7}$$

where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $v \sim \mathcal{N}(0,1)$. This fact is shown in (Maheswaranathan et al., 2019, Section 3.2).

Algorithm 2 Guided zeroth-order method with historical samples

In Algorithm 1, add $p \in \mathbb{N}$ to the input and replace lines 6 and 7 as follows:
$\boldsymbol{s}_k \leftarrow \sum_{j=k-\hat{p}}^{k} \frac{1}{2m_j} \sum_{i=1}^{m_j} (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1}, \boldsymbol{\xi}_j^{1,i}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1}, \boldsymbol{\xi}_j^{2,i}))$,
where $\hat{p} = \min(k, p)$.

3.2. Guided Zeroth-Order Method

We propose Algorithm 1. It begins by probabilistically determining the maximum number R of iterations on Line 1. This stochastic operation is necessary to guarantee the convergence for our method, but practically, the maximum iteration number T can be set as R. Lines 2–15 are the operations of repeatedly updating the decision vector. Lines 3–5 update the decision vector $\boldsymbol{x}_k$ by using the (mini-batch) gradient estimator (4). Lines 6–12 calculate $\boldsymbol{\Sigma}_{k+1}$ by using (5) and (6). Line 13 adjusts the smoothing parameter μ_k: Algorithm 1 starts with a sufficiently large μ_0 and gradually reduces μ_k. It is known that a better local solution can be potentially obtained by making such adjustments to μ_k (Hazan et al., 2016; Iwakiri et al., 2022).[1] Line 14 adjusts the balance parameter α_k. It is updated so that $1-\alpha_{k+1} = \gamma(1-\alpha_k)$, with α_k approaching 1 toward the end. As noted in Section 3.1, this setting promotes a faster decrease in the objective value initially and reduces the bias of the gradient estimate $\boldsymbol{g}_k$ in later iterations.

3.3. Guided Zeroth-Order Method with Historical Samples

Although Algorithm 1 generates efficient random directions, it requires new samples, $\{\boldsymbol{\xi}_k^{3,j}\}_{j=1}^{n_k}$ on line 6 to set $\boldsymbol{\Sigma}$. In order to prevent this increase in sample size, we propose an advanced algorithm that utilizes past samples. Specifically, we change the partial gradient information of (5) as follows.

$$\boldsymbol{s} := \sum_{j=k-\hat{p}}^{k-1} \frac{1}{2m_j} \sum_{i=1}^{m_j} (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\xi}_j^{1,i}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\xi}_j^{2,i})), \tag{8}$$

where $\{\boldsymbol{\xi}_j^{1,i}\}_{i=1}^{m_j}$ and $\{\boldsymbol{\xi}_j^{2,i}\}_{i=1}^{m_j}$ are historical samples from the past j-th iteration. Here, m_j is the batch size of the j-th iteration and $\hat{p} \in \mathbb{N}$ is the window size that determines how far back the past samples are.

In Algorithm 2, the partial gradient information is obtained using samples from past iterations. This allows us to set $\boldsymbol{\Sigma}$ without requiring a new sample.

[1] However, this adjustment is not necessary to guarantee the convergence of our method, and theoretically, we can also use a fixed μ_k.

4. Theoretical Analysis

4.1. Assumptions and Definitions

For the theoretical analysis of our method, we make the following assumptions.

Assumption 4.1. For any $\boldsymbol{x} \in \mathbb{R}^d$, there exists $\sigma \in \mathbb{R}_{\geq 0}$ satisfying

$$\mathbb{E}_{\boldsymbol{\xi} \sim D(\boldsymbol{x})}[(F(\boldsymbol{x}) - f(\boldsymbol{x}, \boldsymbol{\xi}))^2] \leq \sigma^2. \tag{9}$$

Assumption 4.2. For any $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is $L_{\boldsymbol{\xi}}$-Lipschitz continuous with respect to $\boldsymbol{\xi}$. Moreover, for any $\boldsymbol{\xi}$, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is $L_{\boldsymbol{x}}$-Lipschitz continuous with respect to $\boldsymbol{x}$.

Assumption 4.3. For any $\boldsymbol{x} \in \mathbb{R}^d$, there exists $\theta \in \mathbb{R}_{\geq 0}$ satisfying

$$W(D(\boldsymbol{x}), D(\boldsymbol{x}')) \leq \theta \|\boldsymbol{x} - \boldsymbol{x}'\|,$$

where W represents the Wasserstein-1 distance.

Assumption 4.4. $F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi} \sim D(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ is H_F-smooth.

Assumption 4.1 is required for approximating $F(\boldsymbol{x})$ by $f(x, \xi)$ with sample ξ. Since the objective function involves random variables, such an assumption is needed to evaluate the objective value by its sample. The purpose of Assumptions 4.2 and 4.3 is to guarantee the Lipschitz continuity of the objective function (Lemma D.2). It is used to derive the properties of the Gaussian smoothing function (Lemma D.8). Assumption 4.4 is standard in convergence analysis, ensuring the accuracy of the first-order approximation via Taylor expansion. This is because descent methods with (estimated) gradients can be seen as optimizing a first-order approximation of the objective function at each iteration.

Comparison with assumptions in existing studies. Our assumptions are looser than or equal to existing studies (Ray et al., 2022; Hikima & Takeda, 2025b). First, our assumptions are less restrictive than those of (Ray et al., 2022). Assumption 5 in (Ray et al., 2022) implies that $E_{\boldsymbol{\xi} \sim D(\boldsymbol{x})}[(F(\boldsymbol{x}) - f(\boldsymbol{x}, \boldsymbol{\xi}))^2] \leq (2G)^2$ for $G := \sup_{\boldsymbol{x}, \boldsymbol{\xi}} |f(\boldsymbol{x}, \boldsymbol{\xi})|$, which yields our Assumption 4.1. Assumptions 1 and 3 in (Ray et al., 2022) yield our Assumptions 4.2, 4.3, and 4.4. Conversely, we do not require Assumption 1(c) or Assumption 2 in (Ray et al., 2022). Moreover, our assumptions are equivalent to those of (Hikima & Takeda, 2025b). Regarding Assumption 4.4, the existing study (Ray et al., 2022) gives a sufficient condition. It can be found in Appendix C.

Moreover, we define an indicator to evaluate our method.

Definition 4.5 (Sample complexity for ϵ-stationary point)**.** We define the *sample complexity* for an ϵ-stationary point to be the number of samples $\boldsymbol{\xi} \sim D(\boldsymbol{x})$ to obtain $\mathbb{E}[\|\nabla F(\hat{\boldsymbol{x}})\|^2] \leq \epsilon^2$, where $\hat{\boldsymbol{x}}$ is the output of a target algorithm.

As noted in the introduction, sample complexity is an important metric in evaluating algorithms for solving problem (1). This is because, to obtain a sample $\boldsymbol{\xi} \sim D(\boldsymbol{x})$ in practice, a decision $(\boldsymbol{x})$ must be deployed in the real world.

No assumptions about partial gradient information. In our analysis, we do not assume any correlation between partial gradient information (i.e., (5) and (8)) and the true gradient. Therefore, we consider the worst-case scenario for the partial gradient information. Even under this worst-case scenario, our analysis establishes improved sample complexity for our methods compared to the existing method (Hikima & Takeda, 2025b). This improvement is achieved through (i) a reduction in sample complexity by tuning algorithmic parameters such as mini-batch size and step size, and (ii) the incorporation of partial gradient information without increasing the sample complexity, even in the worst-case scenario.

4.2. Sample Complexities of Our Methods

First, we provide a lemma, which indicates a theoretical property of our gradient estimator.

Lemma 4.6. *Suppose that Assumptions 4.1–4.4 hold. For any $\boldsymbol{x} \in \mathbb{R}^d$, the following holds.*

$$\mathbb{E}_{\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})} \left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,i}, \boldsymbol{\xi}^{2,i}\}_{i=1}^m} \left[\left\| \frac{1}{m} \sum_{i=1}^m \boldsymbol{g}(\boldsymbol{x}, \mu, \boldsymbol{u}, \boldsymbol{\xi}^{1,i}, \boldsymbol{\xi}^{2,i}) \right\|^2 \right] \right]$$
$$\leq \frac{24\alpha^2}{d^2}(d+4)^2 \|\nabla F(\boldsymbol{x})\|^2 + 24(1-\alpha)(25-23\alpha)\|\nabla F(\boldsymbol{x})\|^2$$
$$+ 3\mu^2 H_F^2 \left(\frac{16\alpha^3 (d+6)^3}{d^3} + 5488(1-\alpha)^3 \right) + \frac{3\sigma^2}{2\mu^2 m},$$

where $\{\boldsymbol{\xi}^{1,i}\}_{i=1}^m \sim D(\boldsymbol{x} + \mu \boldsymbol{u})$ and $\{\boldsymbol{\xi}^{2,i}\}_{i=1}^m \sim D(\boldsymbol{x} - \mu \boldsymbol{u})$.

Lemma 4.6 gives an upper bound on the second moment of our gradient estimator.[2]

Below, we use Lemmas 3.2 and 4.6 to derive the sample complexity of Algorithm 1.

Theorem 4.7. *Suppose that Assumptions 4.1–4.4 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with $\mu_{\min} = \Theta(\epsilon d^{-\frac{3}{2}})$,*

[2] An existing study (Maheswaranathan et al., 2019) showed the upper-bound of the (normalized) variance of their gradient estimator for given correlation $\rho_i := \frac{\nabla F(\boldsymbol{x})^\top U_i}{\|\nabla F(\boldsymbol{x})\|}$ between partial gradient information (U_i's) and the true gradient ($\nabla F(\boldsymbol{x})$). In our study, we derived an upper bound on the second moment of our gradient estimator (Lemma 4.6) without such ρ_i, leading to a guarantee of convergence for the proposed method. While (Maheswaranathan et al., 2019) demonstrated a trade-off between the unbiasedness and the variance of their gradient estimator by considering ρ_i, our study does not consider ρ_i, and therefore, theoretically, no such trade-off exists in our setting.

$\mu_0 = \Theta(\epsilon d^{-\frac{3}{2}})$ *such that* $\mu_0 \geq \mu_{\min}$, $\beta := \min\left(\frac{\alpha_0 d}{48H_F((d+4)^2+25d^2)}, T^{-\frac{2}{3}}d^{-\frac{1}{3}}\right)$, *and* $\Pr(R=k) := \frac{\beta\alpha_k d^{-1}-24H_F\beta^2\alpha_k^2(d+4)^2d^{-2}-24H_F\beta^2(1-\alpha_k)(25-23\alpha_k)}{\sum_{\ell=0}^{T}(\beta\alpha_\ell d^{-1}-24H_F\beta^2\alpha_\ell^2(d+4)^2d^{-2}-24H_F\beta^2(1-\alpha_\ell)(25-23\alpha_\ell)))}$ *for distribution* D_R. *Then, for any* $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{\Sigma}_0 \in \mathbb{S}^d_{++}$, $\eta \in (0,1]$, $\alpha_0 \in (0,1]$, $\gamma \in [0,1)$, $n_k = O(1)$ *for* $k = 0,\dots T$, $m_k = O(1)$ *for* $k = 0,\dots T$, *the sample complexity required to obtain* $\mathbb{E}[\|\nabla F(\boldsymbol{x}_R)\|^2] \leq \epsilon^2$ *is* $O(\sigma^3\epsilon^{-6}d^4)$.

Moreover, we can derive the sample complexity for Algorithm 2.

Theorem 4.8. *Suppose that Assumptions 4.1–4.4 hold. Let the inputs of Algorithm 1 other than p be as in Theorem 4.7. Then, for any* $p \in \mathbb{N}$, *Algorithm 2 achieves the same sample complexity as in Theorem 4.7.*

The $O(\sigma^3 d^4 \epsilon^{-6})$ sample complexity of our methods has advantages over those of existing methods (Liu et al., 2024a; Hikima & Takeda, 2025b). The sample complexity of the method of (Liu et al., 2024a) is $O(G^6 d^2 \epsilon^{-6})$, where $G = \sup_{\boldsymbol{x},\boldsymbol{\xi}} |f(\boldsymbol{x},\boldsymbol{\xi})|$.[3] Therefore, our methods outperform them when G is large or unbounded.[4] Moreover, the iteration complexity of our method is $O(d^{\frac{1}{2}})$ smaller than that of (Hikima & Takeda, 2025b).

Remark 4.9. To set the inputs of Algorithm 1 according to Theorem 4.7, it is necessary to know the value of H_F in Assumption 4.4. While the existing methods (Ray et al., 2022; Hikima & Takeda, 2025b) also need such information (e.g., γ in (Ray et al., 2022) and σ in (Hikima & Takeda, 2025b)), it may not be known in advance in practice. In such cases, one can begin with a sufficiently large value and refine it using information gathered during the iterations.

5. Experiments

We conducted two experiments on applications of *multiproduct pricing* (Hikima & Takeda, 2025b) and *strategic classification* (Levanon & Rosenfeld, 2021) to show that Algorithms 1 and 2 output solutions with lower objective values compared with the existing methods for the same number of samples. All experiments were conducted on a computer with Intel(R) Xeon(R) CPU E5-2697A v4 (2.60GHz) x2 and 512GB of memory RAM. The program code was implemented in Python 3.12.2.

Compared methods. We implemented the following methods. Details of the parameters can be found in Appendix A.1.2 and A.2.1.

[3]Note that (Liu et al., 2024a) considers a setting in which the sample/data distribution evolves according to a controlled Markov chain, and addresses a more general problem than ours.

[4]Note that even if f is a simple loss function such as squared error, G is unbounded.

GZO-NS. This means the Guided-Zeroth Order method with New Samples, which corresponds to Algorithm 1.
GZO-HS. This means the Guided-Zeroth Order method with Historical Samples, which corresponds to Algorithm 2.
ZO-TG. This is a Zeroth-Order method with a Two-point Gradient estimator. It is consistent with Algorithm 1 where $\boldsymbol{\Sigma}_k = \frac{\boldsymbol{I}_d}{d}$ for all $k \in \{0,\dots,T\}$ and lines 6–12 and line 14 are not executed. It is analogous to the zeroth-order method used in the existing studies ((Hikima & Takeda, 2025b, Algorithm 2) and (Liu et al., 2024a, DFO(0) with $\boldsymbol{g}_{2\mathrm{pt}-\mathrm{II}}$)).
ZO-OG. This is a Zeroth-Order method with a One-point Gradient estimator. It is analogous to the zeroth-order method used in the existing studies (Ray et al., 2022; Liu et al., 2024a).
ZO-OGVR. This is a Zeroth-Order method with a One-point Gradient estimator with a Variance Reduction parameter (Hikima & Takeda, 2025b, Algorithm 1).

5.1. Multiproduct pricing application

We conducted experiments on the same problem as (Hikima & Takeda, 2025b). Specifically, it is the following problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{10}} \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x})}[f(\boldsymbol{x},\boldsymbol{\xi})],$$

where $\boldsymbol{x}$ and $\boldsymbol{\xi}$ denote the price vector and the demand vector for 10 products, respectively. Function f is defined by $f(\boldsymbol{x},\boldsymbol{\xi}) := -s(\boldsymbol{x},\boldsymbol{\xi}) + c(\boldsymbol{\xi})$, where $s(\boldsymbol{x},\boldsymbol{\xi})$ and $c(\boldsymbol{\xi})$ represent the sales and production costs of products, respectively. Distribution $D(\boldsymbol{x})$ represents the (price-dependent) probability distribution that $\boldsymbol{\xi}$ follows. This experiments were semi-synthetic, and some of the parameters were set using real retail data from a supermarket service provider in Japan.[5] Details of how to set each function and distribution are given in Appendix A.1.1.

Setting and Metric

We performed our experiments under the following settings.
Initial points. For all methods, we set the initial points as $\boldsymbol{x}_0 := 0.5 \cdot \boldsymbol{1}$, where $\boldsymbol{1} := (1,\dots,1) \in \mathbb{R}^{10}$.
Metric. To evaluate the output $\hat{\boldsymbol{x}}$, we computed $obj := \frac{1}{10^3}\sum_{q=1}^{10^3} f(\hat{\boldsymbol{x}},\boldsymbol{\xi}^q(\hat{\boldsymbol{x}}))$, where $\boldsymbol{\xi}^q(\hat{\boldsymbol{x}}) \sim D(\hat{\boldsymbol{x}})$.
Termination criteria. We terminated each method once it had taken 5000 samples from $D(\boldsymbol{x})$ for some $\boldsymbol{x}$.

Experimental Results

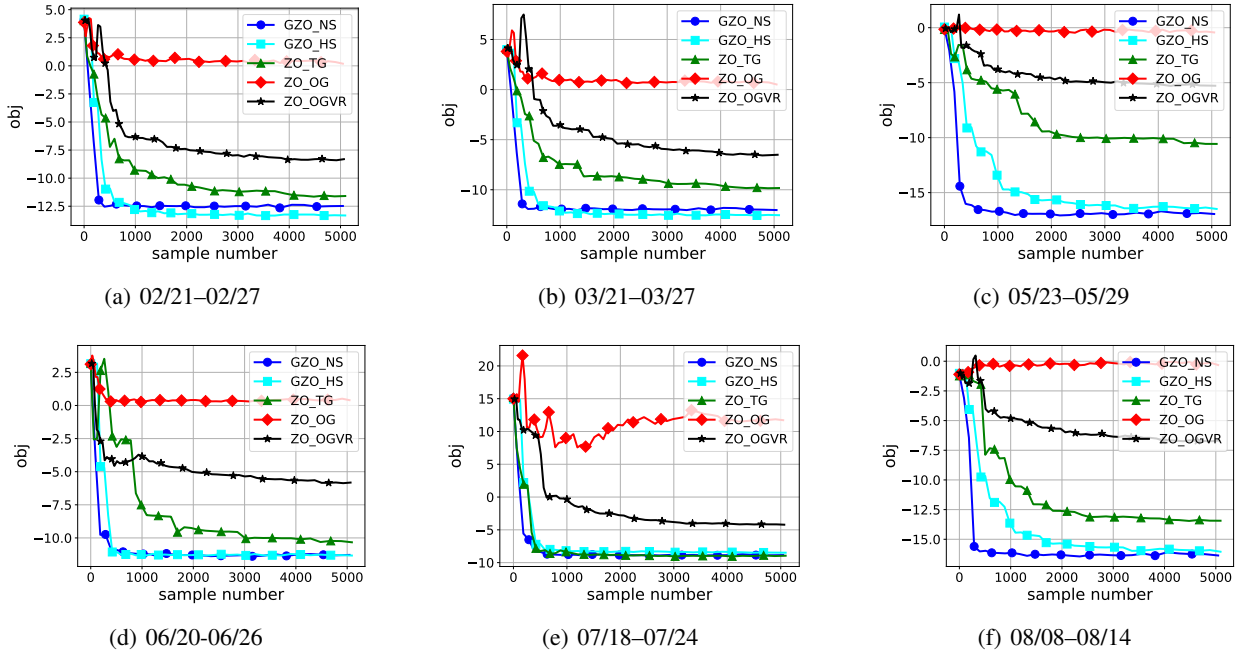
Table 1 shows the results of the experiments using real data from different weeks. Our methods (**GZO-NS** and **GZO-HS**) were superior to the baselines for all weeks of data.

Figure 1 illustrates the objective value (*obj*) obtained with

[5]The data, "New Product Sales Ranking", has been made publicly available by KSP-SP Co., Ltd, http://www.ksp-sp.com. Last accessed on January 28, 2025.

Table 1. Experimental results for multiple product pricing in 20 randomly generated problem instances. The *obj* (*sd*) column represents the average (standard deviation) of the *obj*. The best value of the average *obj* for each experiment is in bold. In all experiments except for 7/18–7/24, the differences in *obj* between our methods (GZO-NS and GZO-HS) and the baselines were statistically significant (two-sided t-test: $p < 0.05$). For 7/18–7/24, the difference between GZO-HS and ZO-TG was not statistically significant, while the difference between GZO-NS and the baselines remained significant.

date	GZO-NS		GZO-HS		ZO-TG		ZO-OG		ZO-OGVR	
	obj	sd	obj	sd	obj	sd	obj	sd	obj	sd
2/21–2/27	-12.53	0.99	**-12.85**	0.70	-10.92	1.21	5.10	10.33	-6.30	1.75
3/21–3/27	-12.52	1.29	**-12.58**	1.09	-10.52	1.45	1.56	3.38	-6.41	1.64
5/23–5/29	**-15.70**	1.04	-15.41	1.17	-10.31	2.29	-0.06	1.59	-5.60	2.07
6/20–6/26	-9.06	1.30	**-9.52**	0.69	-8.81	0.82	10.15	17.28	-5.65	1.96
7/18–7/24	**-8.43**	0.99	-8.38	1.29	-7.98	1.42	10.38	19.72	-3.47	2.43
8/08–8/14	**-14.98**	1.01	-14.93	1.28	-10.30	2.41	-0.53	1.87	-6.11	1.55



(a) 02/21–02/27 (b) 03/21–03/27 (c) 05/23–05/29

(d) 06/20-06/26 (e) 07/18–07/24 (f) 08/08–08/14

Figure 1. Change in *obj* in the experiments on multiproduct pricing. Each graph shows the result for one problem instance for each week. The horizontal axis indicates the cumulative number of samples from $D(\boldsymbol{x})$, and the vertical axis indicates *obj*.

each method against the cumulative number of samples from $D(\boldsymbol{x})$ in its optimization process. The figure implies that our methods reduce the objective value more stably than the baselines. This is due to the following reasons: (i) **ZO-TG** and **ZO-OGVR** could not efficiently decrease the objective value because they determined the updating direction $\boldsymbol{u}_k$ according to the standard Gaussian distribution without using information on the function f; (ii) **ZO-OG** could not stably decrease the objective value because of the large variance of its gradient estimator.[6]

[6]The large variance of the one-point gradient estimator used by ZO-OG is also noted in (Hikima & Takeda, 2025b).

5.2. Strategic Classification with Unknown Agents' Cost Functions

We conducted experiments on the application of strategic classification with a real dataset from (Yeh & hui Lien, 2009).[7] We considered a variant of (Levanon & Rosenfeld, 2021, Section 4) where the decision maker wants to optimize the parameter $\boldsymbol{x}$ of the classifier to decide whether to provide loan financing for a strategic agent with feature $\boldsymbol{\xi}_F \in \mathbb{R}^{11}$ and label $L \in \{0, 1\}$. Here, $\boldsymbol{\xi}_F$ includes credit-card spending patterns, while L indicates whether the agent

[7]As with (Levanon & Rosenfeld, 2021), we used a processed version of the data from (Ustun et al., 2019).

defaults on a payment. Since the agent can vary their own features according to the parameter $\boldsymbol{x}$, $(\boldsymbol{\xi}_F, L)$ follows a decision-dependent distribution $D(\boldsymbol{x})$. Here, we used the following loss function:

$$f(\boldsymbol{\xi}_F, L; \boldsymbol{x}) := L \log\left(\frac{1}{1+e^{-(\boldsymbol{x}_{[11]}^\top \boldsymbol{\xi} + x_{12})}}\right) + (1-L)\log\left(1 - \frac{1}{1+e^{-(\boldsymbol{x}_{[11]}^\top \boldsymbol{\xi} + x_{12})}}\right),$$

where $\boldsymbol{x}_{[11]} := (x_1, x_2, \ldots, x_{11})$.

The loss-minimization problem is as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{12}} \quad \mathbb{E}_{(\boldsymbol{\xi}_F, L)\sim D(\boldsymbol{x})}\left[f(\boldsymbol{\xi}_F, L; \boldsymbol{x})\right].$$

Settings of unknown $D(\boldsymbol{x})$. We assumed that each agent has the true feature $\boldsymbol{\xi}_{true} \sim D_{true}$ but may change its features according to the parameter $\boldsymbol{x}$ of the classifier and the cost associated with altering its feature. Specifically, each agent modifies its features by using the following response mapping:

$$\boldsymbol{\xi}_F := \arg\max_{\boldsymbol{\xi}} r(\boldsymbol{\xi}, \boldsymbol{x}) - c(\boldsymbol{\xi}, \boldsymbol{\xi}_{true}, \tau),$$

where

$$r(\boldsymbol{\xi}, \boldsymbol{x}) := \begin{cases} 2, & \text{if } \boldsymbol{x}_{[11]}^\top \boldsymbol{\xi} + x_{12} \geq 0, \\ 0, & \text{if } \boldsymbol{x}_{[11]}^\top \boldsymbol{\xi} + x_{12} < 0, \end{cases}$$
$$c(\boldsymbol{\xi}, \boldsymbol{\xi}_{true}, \tau) := \tau \|\boldsymbol{\xi} - \boldsymbol{\xi}_{true}\|^2.$$

Here, $r(\boldsymbol{\xi}, \boldsymbol{x})$ represents the profit for each agent: if $\boldsymbol{x}_{[11]}^\top \boldsymbol{\xi} + x_{12} \geq 0$, then the agent receives a positive classification and is rewarded.[8] Function c represents the cost of modifying their features: changing one's own features incurs a cost based on the distance of the modified feature vector from the true feature vector. Here, $\tau \in \mathbb{R}_{>0}$ is a constant that controls the magnitude of the cost. We consider the set of real data to be the set of $(\boldsymbol{\xi}_{true}, L)$ and set $D(\boldsymbol{x})$ with the above settings. Note that the information on $D(x)$ was not used by any of the methods.

Remark 5.1. This setting does not satisfy the smoothness in Assumption 4.3. This is because the data distribution changes discontinuously at the point where the cost exceeds the gain for some agents. Despite this unfavorable setting for our methods, the experimental results described later show that they still perform well numerically.

SETTING AND METRIC

We performed our experiments under the following settings.
Initial points. For all methods, we set the initial points as $\boldsymbol{x}_0 := \mathbf{1}$, where $\mathbf{1} := (1, \ldots, 1) \in \mathbb{R}^{12}$.
Metrics. For the output of each method, we used the training loss, the test loss, the test AUC, and the test accuracy as metrics.
Termination criteria. We terminated each method once it had taken 10000 samples from $D(\boldsymbol{x})$ for some $\boldsymbol{x}$.

EXPERIMENTAL RESULTS

Table 2 shows the results of the experiments with different τ.[9] Note that $\tau \in \mathbb{R}_{>0}$ is a constant that controls the magnitude of the cost. If $\tau \in \mathbb{R}_{>0}$ is small, distribution $D(\boldsymbol{x})$ is more likely to vary with decision vector $\boldsymbol{x}$. The results show that our methods (**GZO-NS** and **GZO-HS**) were superior to the baselines for all τ. Moreover, the performance of our methods is less affected by changes in τ than existing methods, which indicates our methods are less sensitive to the changeability of distribution $D(\boldsymbol{x})$.

6. Conclusion

We proposed two zeroth-order methods for solving non-convex stochastic problems with decision-dependent distributions; they utilize partial gradient information derived from the known structure of f. The first method obtains partial gradient information from samples at the current iteration, and the second one obtains partial gradient information from historical samples. Our theoretical analysis showed that they converged to stationary points and provided the worst-case sample complexity. Our experimental results showed that our methods outperformed the conventional zeroth-order methods.

Future work includes exploring efficient methods for adjusting the balance parameter α, which controls the weight of the partial gradient information. We used the first term in (3) as partial gradient information, whose importance is affected by the size of the second term. Therefore, we can adjust α according to the size of the second term. Similarly, it is expected that a method for determining the range p of past samples to be referenced in Algorithm 2 can also be derived. If the reliability of past samples can be theoretically established, the value of p can be accordingly set.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

[8] The reward value of 2 is set based on the existing study (Levanon & Rosenfeld, 2021).

[9] Note that the performance of each method in our experiments is lower than that in typical experiments without strategic agents because the strategic agent manipulates its own features to cheat the classifier.

Table 2. Results of experiments on strategic classification for 20 randomly generated problem instances. The *metric* (*sd*) column represents its average value (standard deviation). The best value of each metric for each experiment is in bold. In all experiments, the differences in each metric between our methods (GZO-NS and GZO-HS) and the baselines are significant (two-sided t-test: $p < 0.05$).

τ	GZO-NS		GZO-HS		ZO-TG		ZO-OG		ZO-OGVR	
	train loss	sd	train loss	sd	train loss	sd	train loss	sd	train loss	sd
0.5	0.87	0.04	**0.83**	0.03	1.21	0.08	2.75	1.80	1.51	0.62
1.0	0.88	0.05	**0.84**	0.03	1.27	0.08	2.98	1.83	1.63	0.63
2.0	0.86	0.05	**0.81**	0.03	1.34	0.10	3.44	1.92	1.83	0.73
4.0	0.84	0.05	**0.79**	0.03	1.35	0.12	3.96	2.09	2.06	0.93
8.0	0.83	0.04	**0.78**	0.03	1.35	0.12	4.39	2.27	2.24	1.09

τ	GZO-NS		GZO-HS		ZO-TG		ZO-OG		ZO-OGVR	
	test loss	sd	test loss	sd	test loss	sd	test loss	sd	test loss	sd
0.5	0.93	0.04	**0.89**	0.03	1.31	0.09	2.89	1.91	1.63	0.67
1.0	0.96	0.05	**0.91**	0.04	1.38	0.09	3.12	1.94	1.75	0.69
2.0	0.93	0.06	**0.87**	0.04	1.43	0.11	3.55	2.03	1.96	0.79
4.0	0.90	0.05	**0.84**	0.04	1.45	0.13	4.04	2.20	2.20	1.01
8.0	0.88	0.05	**0.83**	0.04	1.45	0.14	4.44	2.39	2.38	1.18

τ	GZO-NS		GZO-HS		ZO-TG		ZO-OG		ZO-OGVR	
	test AUC	sd	test AUC	sd	test AUC	sd	test AUC	sd	test AUC	sd
0.5	**0.62**	0.03	**0.62**	0.02	0.46	0.03	0.48	0.12	0.44	0.08
1.0	**0.62**	0.03	**0.62**	0.02	0.47	0.03	0.49	0.13	0.45	0.09
2.0	**0.66**	0.03	**0.66**	0.02	0.48	0.04	0.50	0.14	0.46	0.10
4.0	0.67	0.03	**0.68**	0.02	0.49	0.04	0.50	0.15	0.46	0.11
8.0	0.68	0.02	**0.69**	0.02	0.50	0.05	0.51	0.15	0.46	0.12

τ	GZO-NS		GZO-HS		ZO-TG		ZO-OG		ZO-OGVR	
	test acc	sd	test acc	sd	test acc	sd	test acc	sd	test acc	sd
0.5	**0.55**	0.03	0.53	0.03	0.47	0.02	0.49	0.06	0.47	0.04
1.0	**0.57**	0.03	0.56	0.03	0.47	0.03	0.51	0.08	0.47	0.05
2.0	**0.60**	0.03	0.59	0.03	0.48	0.02	0.51	0.10	0.48	0.06
4.0	**0.62**	0.03	**0.62**	0.03	0.47	0.03	0.51	0.10	0.48	0.07
8.0	**0.63**	0.03	**0.63**	0.02	0.49	0.03	0.52	0.11	0.48	0.08

References

Bergstra, J. and Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 2012.

Chen, Y., Tang, W., Ho, C.-J., and Liu, Y. Performative prediction with bandit feedback: learning through reparameterization. In *International Conference on Machine Learning*, pp. 7298 – 7324, 2024.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811.*, 2018.

Freund, J. E. and Walpole, R. E. *Mathematical Statistics*. Prentice-Hall, Inc., 1986.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. On graduated optimization for stochastic non-convex problems. In *International Conference on Machine Learning*, pp. 1833–1841, 2016.

Hikima, Y. and Takeda, A. Stochastic approach for price optimization problems with decision-dependent uncertainty. *European Journal of Operational Research*, 322 (2):541–553, 2025a.

Hikima, Y. and Takeda, A. Zeroth-order methods for non-convex stochastic problems with decision-dependent dis-

tributions. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 17195–17203, 2025b.

Iwakiri, H., Wang, Y., Ito, S., and Takeda, A. Single loop gaussian homotopy method for non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7065–7076, 2022.

Izzo, Z., Ying, L., and Zou, J. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pp. 4641–4650, 2021.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pp. 9760–9785, 2022.

Khorsandi, P., Gupta, R., Mofakhami, M., Lacoste-Julien, S., and Gidel, G. Tight lower bounds and improved convergence in performative prediction. In *OPT 2024: Optimization for Machine Learning*, 2024.

Levanon, S. and Rosenfeld, N. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253, 2021.

Li, Q. and Wai, H.-T. Stochastic optimization schemes for performative prediction with nonconvex loss. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8673–8697, 2024.

Lin, L. and Zrnic, T. Plug-in performative optimization. In *International Conference on Machine Learning*, pp. 30546 – 30565, 2024.

Liu, H., Li, Q., and Wai, H.-T. Two-timescale derivative free optimization for performative prediction with markovian data. In *International Conference on Machine Learning*, pp. 31425–31450, 2024a.

Liu, T., Lin, Y., and Zhou, E. Bayesian stochastic gradient descent for stochastic optimization with streaming input data. *SIAM Journal on Optimization*, 34(1):389–418, 2024b.

Maheswaranathan, N., Metz, L., Tucker, G., Choi, D., and Sohl-Dickstein, J. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *International Conference on Machine Learning*, pp. 4264–4273, 2019.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939, 2020.

Miller, J. P., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720, 2021.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093, 2023.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609, 2020.

Petersen, K. B., Pedersen, M. S., et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Powell, M. J. D. The NEWUOA software for unconstrained optimization without derivatives. In *Large-Scale Nonlinear Optimization*, pp. 255–297. Springer US, 2006.

Ragonneau, T. M. and Zhang, Z. PDFO: a cross-platform package for powell's derivative-free optimization solvers. *Mathematical Programming Computation*, 16(4):535–559, 2024.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 8081–8088, 2022.

Roy, A., Balasubramanian, K., and Ghadimi, S. Constrained stochastic nonconvex optimization with state-dependent markov data. In *Advances in neural information processing systems*, volume 35, pp. 23256–23270, 2022.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.

Xue, J. and Shen, B. A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science & Control Engineering*, 8(1):22–34, 2020.

Yeh, I.-C. and hui Lien, C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009.

A. Details of Our Experiments

A.1. Experiments on multiproduct pricing

A.1.1. PROBLEM SETTING

We performed semi-synthetic experiments based on (Hikima & Takeda, 2025b). Here, some of the parameters were set using real retail data from a supermarket service provider in Japan.[10] The experiments simulated a scenario in which a seller determines the prices of multiple products ($n = 10$) for multiple buyers ($m = 40$). Each buyer purchases at most one unit of any product. Let $\boldsymbol{x} := (x_1, x_2, \ldots, x_{10}) \in \mathbb{R}^{10}$ be the price vector (decision vector) for the products. Let $\boldsymbol{\xi} \in \{0, 1, \ldots, m\}^{11}$ denote the stochastic demand vector of the products, where ξ_i for $i = 1, \ldots, 10$ is the number of sales of each product, and ξ_{11} is the number of buyers who do not purchase any product.

Then, the objective is to solve the following revenue-maximization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^{10}} \mathbb{E}_{\boldsymbol{\xi} \sim D(\boldsymbol{x})} \left[f(\boldsymbol{x}, \boldsymbol{\xi}) \right],$$

where $f(\boldsymbol{x}, \boldsymbol{\xi}) := -s(\boldsymbol{x}, \boldsymbol{\xi}) + c(\boldsymbol{\xi})$ and $D(\boldsymbol{x})$ is the probability distribution of demand $\boldsymbol{\xi}$. Here, $s(\boldsymbol{x}, \boldsymbol{\xi})$ and $c(\boldsymbol{\xi})$ are the sales and production cost function, respectively. They are defined as:

$$s(\boldsymbol{x}, \boldsymbol{\xi}) := \sum_{i=1}^{n} x_i \xi_i, \quad c(\boldsymbol{\xi}) := \sum_{i=1}^{n} c_i(\xi_i),$$

where

$$c_i(\xi_i) := \begin{cases} 2w_i \xi_i, & \xi_i \leq l_i, \\ w_i(\xi_i - l_i) + 2w_i l_i, & l_i < \xi_i \leq u_i, \\ 3w_i(\xi_i - u_i) + w_i(u_i - l_i) + 2w_i l_i, & \xi_i > u_i. \end{cases}$$

Here, we set $l_i := \frac{0.5m}{n}$, $u_i := \frac{1.5m}{n}$, and $w_i := \rho_i \theta_i$, where ρ_i is a random variable generated from a uniform distribution of $[0.25, 0.5]$ and θ_i is the normalized recorded average selling price for each product i. The function c_i reflects the scenario where the production cost rate varies based on the number of units sold.

Settings of unknown $D(\boldsymbol{x})$. We assume that buyers choose one product stochastically. Each buyer chooses product $i \in I := \{1, \ldots, n\}$ with probability $p_i(\boldsymbol{x}) = \frac{e^{\gamma_i(\theta_i - x_i)}}{a_0 + \sum_{j=1}^{n} e^{\gamma_j(\theta_j - x_j)}}$ or does not choose any product with probability $p_0(\boldsymbol{x}) = \frac{a_0}{a_0 + \sum_{j=1}^{n} e^{\gamma_j(\theta_j - x_j)}}$. Here, we let $\gamma_i := \frac{2\pi}{\sqrt{6}\theta_i}$ and let $a_0 := 0.1n$. Then, $\Pr(\xi_i \mid \boldsymbol{x}) = \binom{m}{\xi_i} p_i(\boldsymbol{x})^{\xi_i}$. Note that the information on $D(x)$ was not used by any of the methods.

A.1.2. PARAMETERS OF METHODS.

GZO-NS. This is Algorithm 1 with $\mu_0 := 0.2$, $\mu_{\min} := 0.0001$, $\alpha_0 = 0$, $\beta_k := 0.01 \cdot 0.95^k$, $\eta = 0.95$, $\gamma = 0.98$, $m_k := 30 + 2k$, and $n_k := 30 + 2k$, where k is the current iteration number.
GZO-HS. This is Algorithm 2 with $p = 1$; the other parameters are the same as in GZO-NS.
ZO-TG. This is consistent with Algorithm 1 where $\boldsymbol{\Sigma}_k = \frac{\boldsymbol{I}_d}{d}$ for all $k \in \{0, \ldots, T\}$ and lines 6–12 and line 14 are not executed. The other parameters are the same as in GZO-NS.
ZO-OG. This is consistent with Algorithm 1 where $\boldsymbol{\Sigma}_k = \boldsymbol{I}_d$, $\boldsymbol{g}_k = \frac{1}{m_k} \sum_{j=1}^{m_k} \frac{f(\boldsymbol{x}_k + \mu_k \boldsymbol{u}_k, \boldsymbol{\xi}_k^{1,j})}{\mu_k} \boldsymbol{u}_k$ for all $k \in \{0, \ldots, T\}$, and lines 6–12 and line 14 are not executed. We let $\beta := 10^{-5}$ and $\mu_k := 0.001$; the other parameters are the same as in GZO-NS.[11]
ZO-OGVR. This is consistent with (Hikima & Takeda, 2025b, Algorithm 1). We let $\beta := 0.001 \cdot 0.95^k$, $c_0 := \sum_{j=1}^{20} f(\boldsymbol{x}_0, \boldsymbol{\xi}^j(\boldsymbol{x}_0))$, $s_{\max} := 10$, and $M = 0.1$; the other parameters are the same as in GZO-NS.

[10] The data, "New Product Sales Ranking", has been made publicly available by KSP-SP Co., Ltd, http://www.ksp-sp.com. Last accessed on January 28, 2025.

[11] The step size of **ZO-OG** is set small compared with those of other methods. This is because the variance of the one-point gradient estimator is large, which causes the objective values to diverge.

A.1.3. DATA DETAILS.

The experimental data, "New Product Sales Ranking", provided by KSP-SP Co., Ltd, includes confectionery price data. We used the actual prices of confectionery as θ in the buyer's probability function for purchase. We also determined the cost of each product based on the actual prices.

A.2. Experiments on Strategic Classification

A.2.1. PARAMETERS OF METHODS.

GZO-NS. This is Algorithm 1 with $\mu_0 := 10$, $\mu_{\min} := 0.1$, $\alpha_0 = 0$, $\beta := 0.95^k$, $\eta := 0.95$, $\gamma := 0.98$, $m_k := 30 + 2k$, and $n_k := 100 + 2k$, where k is the current iteration number.
GZO-HS. This is Algorithm 2 with $p := 1$; the other parameters are the same as in GZO-NS.
ZO-TG. This is consistent with Algorithm 1 where $\boldsymbol{\Sigma}_k := \frac{\boldsymbol{I}_d}{d}$ for all $k \in \{0, \ldots, T\}$; the other parameters are the same as in GZO-NS.
ZO-OG. This is consistent with Algorithm 1 where $\boldsymbol{\Sigma}_k = \boldsymbol{I}_d$ and $\boldsymbol{g}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} \frac{f(\boldsymbol{x}_k+\mu_k \boldsymbol{u}_k, \boldsymbol{\xi}_k^{1,j})}{\mu_k}\boldsymbol{u}_k$ for all $k \in \{0, \ldots, T\}$. We let $\beta := 0.05$ and $\mu_k := 0.1$; the other parameters are the same as in GZO-NS.
ZO-OGVR. This is consistent with (Hikima & Takeda, 2025b, Algorithm 1). We let $\beta := d^{-\frac{1}{2}} \cdot 0.95^k$, $c_0 := \sum_{j=1}^{20} f(\boldsymbol{x}_0, \boldsymbol{\xi}^j(\boldsymbol{x}_0))$, $s_{\max} := 10$, and $M = 0.1$; the other parameters are the same as in GZO-NS.

A.2.2. DATA DETAILS.

The experimental dataset (Yeh & hui Lien, 2009) includes features describing credit-card spending patterns, along with labels indicating default on payment. As with (Levanon & Rosenfeld, 2021), we used the preprocessed version of the data by (Ustun et al., 2019). The dataset includes $n = 11$ features. We divided 13272 data samples into a 12272-sample training set and 1000-sample test set in our experiments.

B. Additional Experiments

B.1. Statistical information on objective values during iteration

Since Tables 1 and 2 show the statistical results at the final iteration, we give the figures to show averaged curves and error bars across multiple instances. Figure 2 shows the average and standard deviation of the objective values across iterations, obtained from experiments using data from each week in the multi-product pricing application. These results indicate that our methods are superior to existing methods even during iterations.

B.2. Statistical information on gradient norms during iteration

Figure 3 shows the mean and standard deviation of the gradient norm in the application of multi-product pricing. This graph shows that the proposed method reduces the gradient norm with fewer samples than existing methods.

C. Sufficient condition for Assumption 4.4

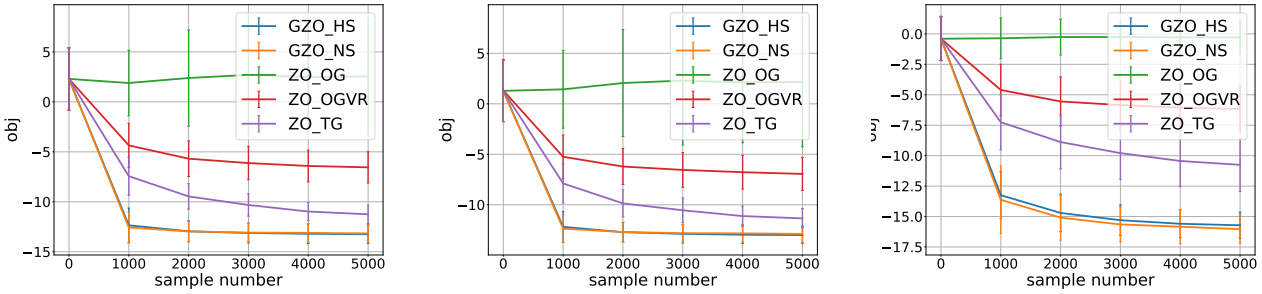
Lemma C.1. *(Ray et al., 2022, Lemma 1) Suppose that there exist a matrix $\boldsymbol{A}$ and distribution D' such that*

$$\boldsymbol{\xi} \sim D(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{\xi} = \boldsymbol{\nu} + \boldsymbol{A}\boldsymbol{x},$$

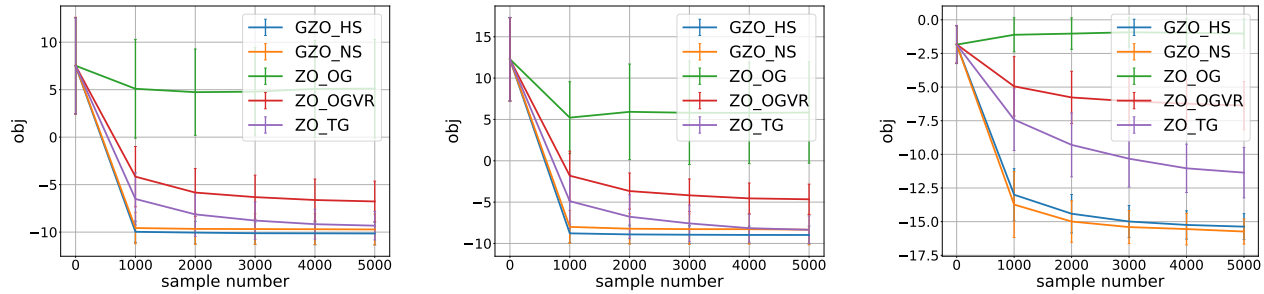
where $\boldsymbol{\nu} \sim D'$. Moreover, suppose that $f(\boldsymbol{x}, \boldsymbol{\xi})$ is ρ-smooth with respect to both $\boldsymbol{x}$ and $\boldsymbol{\xi}$. Then, Assumption 4.4 holds with

$$H_F := \sqrt{\rho^2(1 + \|A\|_{op}^2)\max(1, \|A\|_{op}^4)},$$

where $\|A\|_{op}$ is the operator norm of A, that is, $\|A\|_{op} := \sup_{x \in \mathbb{R}^d, x \neq 0} \frac{\|Ax\|}{\|x\|}$.

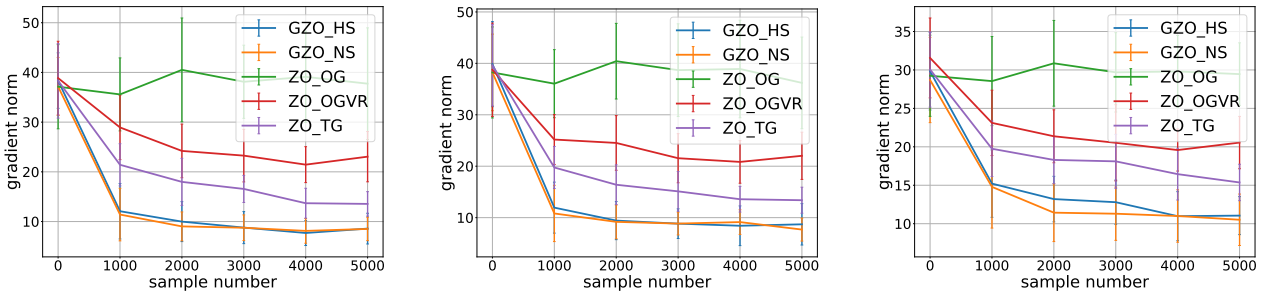


(a) 02/21–02/27 (b) 03/21–03/27 (c) 05/23–05/29

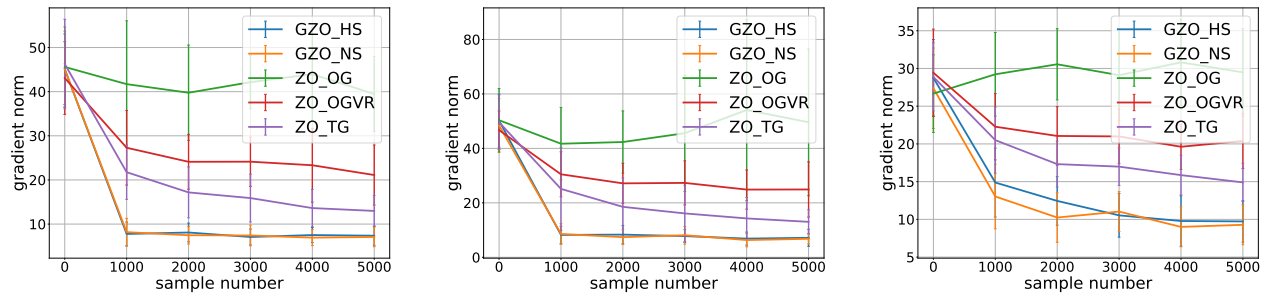


(d) 06/20-06/26 (e) 07/18–07/24 (f) 08/08–08/14

Figure 2. Changes in the mean and standard deviation of *obj* for 20 problem instances relative to the number of samples in the experiments on multiproduct pricing. Each graph shows the result for each week.



(a) 02/21–02/27 (b) 03/21–03/27 (c) 05/23–05/29



(d) 06/20-06/26 (e) 07/18–07/24 (f) 08/08–08/14

Figure 3. Changes in the mean and standard deviation of the approximated gradient norm of the objective function (calculated by (3) with 1000 samples) for 20 problem instances relative to the number of samples in the experiments on multiproduct pricing. Each graph shows the result for each week.

D. Proofs

D.1. Technical lemmas

Here, we provide technical lemmas that are needed to prove the lemmas and theorems in our paper. Technical Lemmas D.1–D.4 are from existing research, while Technical Lemmas D.5–D.12 are newly proved by us.

Technical Lemma D.1. (Nesterov & Spokoiny, 2017, Lemma 1)

For $p \in [0, 2]$,

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\|\boldsymbol{w}\|^p\right] \leq d^{\frac{p}{2}}. \tag{10}$$

For $p \geq 2$,

$$\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\|\boldsymbol{w}\|^p\right] \leq (d+p)^{\frac{p}{2}}. \tag{11}$$

Technical Lemma D.2. (Jagadeesan et al., 2022, Lemma 2.1)

Under Assumptions 4.2 and 4.3, function $F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ is L_F-Lipschitz, where $L_F := L_{\boldsymbol{x}} + \theta L_{\boldsymbol{\xi}}$.

Technical Lemma D.3. (Freund & Walpole, 1986, p183)

For any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m\sim D(\boldsymbol{x})}\left[\left\|\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}, \boldsymbol{\xi}^j) - \mathbb{E}_{\boldsymbol{\xi}\sim D(\boldsymbol{x})}[f(\boldsymbol{x}, \boldsymbol{\xi})]\right\|^2\right] \leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\xi}'\sim D}\left[\|f(\boldsymbol{x}, \boldsymbol{\xi}') - \mathbb{E}_{\boldsymbol{\xi}\sim D}[f(\boldsymbol{x}, \boldsymbol{\xi})]\|^2\right].$$

Technical Lemma D.4. (Maheswaranathan et al., 2019, Section 3.2)

Let $\boldsymbol{\Sigma} := \alpha\frac{\boldsymbol{I}_d}{d} + (1-\alpha)\boldsymbol{h}\boldsymbol{h}^\top$ for $d \in \mathbb{N}$, $\boldsymbol{h} \in \mathbb{R}^d$, and $\alpha \in (0, 1]$. Let $\boldsymbol{u} = \sqrt{\frac{\alpha}{d}}\boldsymbol{w} + \sqrt{1-\alpha}\boldsymbol{h}v$, where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $v \sim \mathcal{N}(0, 1)$. Then, the distribution of $\boldsymbol{u}$ is $\mathcal{N}(0, \boldsymbol{\Sigma})$.

Technical Lemma D.5. Let $\boldsymbol{\Sigma} := \alpha\frac{\boldsymbol{I}_d}{d} + (1-\alpha)\boldsymbol{h}\boldsymbol{h}^\top$ for $d \in \mathbb{N}$, $\boldsymbol{h} \in \mathbb{R}^d$ such that $\|\boldsymbol{h}\| = 1$ or $\boldsymbol{h} = \boldsymbol{0}$, and $\alpha \in (0, 1]$. Then,

$$\frac{\alpha}{d} \leq \|\boldsymbol{\Sigma}\| \leq \frac{\alpha}{d} + 1 - \alpha,$$

where $\|\boldsymbol{\Sigma}\|$ is the spectral norm of $\boldsymbol{\Sigma}$. Moreover,

$$\|\boldsymbol{\Sigma}^{-1}\| = \frac{d}{\alpha}.$$

Proof. Since $\boldsymbol{\Sigma}$ is a positive semi-definite matrix, its spectral norm is equal to its largest eigenvalue. Let λ be an eigenvalue of $\boldsymbol{\Sigma}$. Then, there exists an eigenvector $\boldsymbol{x} \neq \boldsymbol{0}$ that satisfies

$$\lambda\boldsymbol{x} = \boldsymbol{\Sigma}\boldsymbol{x} = \frac{\alpha}{d}\boldsymbol{x} + (1-\alpha)\boldsymbol{h}^\top\boldsymbol{x}\boldsymbol{h}.$$

From the above equation, $\lambda = \frac{\alpha}{d}$ when $\alpha = 1$. When $\alpha = 0$, $\boldsymbol{x}$ satisfies $\boldsymbol{x} = c\boldsymbol{h}$ for some constant $c \in \mathbb{R}$ or satisfies $\boldsymbol{h}^\top\boldsymbol{x} = 0$. In the case that $\boldsymbol{x} = c\boldsymbol{h}$, we have $\boldsymbol{h} \neq 0$ since $\boldsymbol{x} \neq 0$. Then, we have $\lambda = \frac{\alpha}{d} + (1-\alpha)$ since $\|\boldsymbol{h}\| = 1$ from the assumption. In the case that $\boldsymbol{h}^\top\boldsymbol{x} = 0$, we have $\lambda = \frac{\alpha}{d}$. Therefore, from the fact that $\alpha \in (0, 1]$, we have $\frac{\alpha}{d} \leq \|\boldsymbol{\Sigma}\| \leq \frac{\alpha}{d} + 1 - \alpha$.

Next, we have that $\|\boldsymbol{\Sigma}^{-1}\| = \frac{1}{\boldsymbol{\Sigma}_{\min}}$, where $\boldsymbol{\Sigma}_{\min}$ is the smallest singular value of the matrix $\boldsymbol{\Sigma}$. Since the eigenvalues of the positive-definite matrix $\boldsymbol{\Sigma}$ are equal to its singular values, we have that $\|\boldsymbol{\Sigma}^{-1}\| = \frac{1}{\lambda_{\min}}$, where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$. From the previous discussion, $\lambda_{\min} = \frac{\alpha}{d}$. Therefore, $\|\boldsymbol{\Sigma}^{-1}\| = \frac{d}{\alpha}$. □

Technical Lemma D.6. For any $n \in \mathbb{N}$, $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{y} \in \mathbb{R}^d$,

$$\|\boldsymbol{x} + \boldsymbol{y}\|^n \leq 2^{n-1}\|\boldsymbol{x}\|^n + 2^{n-1}\|\boldsymbol{y}\|^n.$$

Proof. Lemma D.6 holds if

$$\left\|\frac{\boldsymbol{x}+\boldsymbol{y}}{2}\right\|^n \leq \frac{\|\boldsymbol{x}\|^n+\|\boldsymbol{y}\|^n}{2}.$$

Due to the symmetry between $\boldsymbol{x}$ and $\boldsymbol{y}$, we can assume that $\|\boldsymbol{x}\| \geq \|\boldsymbol{y}\|$ without loss of generality. Let $\boldsymbol{z} := \frac{\boldsymbol{x}}{\|\boldsymbol{y}\|}$. Then, $\|\boldsymbol{z}\| \geq 1$. Moreover,

$$\left\|\frac{\boldsymbol{x}+\boldsymbol{y}}{2}\right\|^n = \left\|\frac{\boldsymbol{z}+\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}}{2}\right\|^n \|\boldsymbol{y}\|^n \leq \left(\frac{\|\boldsymbol{z}\|+\frac{\|\boldsymbol{y}\|}{\|\boldsymbol{y}\|}}{2}\right)^n \|\boldsymbol{y}\|^n = \left(\frac{\|\boldsymbol{z}\|+1}{2}\right)^n \|\boldsymbol{y}\|^n.$$

Therefore, Lemma D.6 holds if $\left(\frac{\|\boldsymbol{z}\|+1}{2}\right)^n \|\boldsymbol{y}\|^n \leq \frac{\|\boldsymbol{x}\|^n+\|\boldsymbol{y}\|^n}{2}$, that is, $\left(\frac{\|\boldsymbol{z}\|+1}{2}\right)^n \leq \frac{\|\boldsymbol{z}\|^n+1}{2}$ for $\|\boldsymbol{z}\| \geq 1$. Here, let $f(s) := \frac{s^n+1}{2} - \left(\frac{s+1}{2}\right)^n$. Then, for $s \geq 1$,

$$f'(s) = \frac{ns^{n-1}}{2} - \frac{n(s+1)^{n-1}}{2^n} = \frac{n}{2}\left(s^{n-1} - \left(\frac{s+1}{2}\right)^{n-1}\right) \geq 0.$$

Since $f(1) = 0$, we have $f(s) \geq f(1) = 0$ for $s \geq 1$. Therefore, $\left(\frac{\|\boldsymbol{z}\|+1}{2}\right)^n \leq \frac{\|\boldsymbol{z}\|^n+1}{2}$ for $\|\boldsymbol{z}\| \geq 1$. □

Technical Lemma D.7. Let $\boldsymbol{\Sigma} := \alpha\frac{\boldsymbol{I}_d}{d} + (1-\alpha)\boldsymbol{h}\boldsymbol{h}^\top$ for $d \in \mathbb{N}$, $\boldsymbol{h} \in \mathbb{R}^d$ such that $\|\boldsymbol{h}\| = 1$ or $\boldsymbol{h} = \boldsymbol{0}$, and $\alpha \in (0, 1]$. Then,

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^2] \leq 1, \tag{12}$$

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^6] \leq \frac{32\alpha^3(d+6)^3}{d^3} + 10976(1-\alpha)^3. \tag{13}$$

Proof. Since $\|\boldsymbol{u}\|^2 = \mathrm{tr}(\boldsymbol{u}\boldsymbol{u}^\top)$ from (Petersen et al., 2008, (17)), we obtain that

$$\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^2] = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\mathrm{tr}(\boldsymbol{u}\boldsymbol{u}^\top)] = \mathrm{tr}(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\boldsymbol{u}\boldsymbol{u}^\top]) = \mathrm{tr}(\boldsymbol{0}\boldsymbol{0}^\top + \boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^d \lambda_i,$$

where λ_i's are the eigenvalues of $\boldsymbol{\Sigma}$. Therefore, (12) holds if $\sum_{i=1}^d \lambda_i \leq 1$. Let $\boldsymbol{x} \neq \boldsymbol{0}$ be an eigenvector of $\boldsymbol{\Sigma}$. Then,

$$\lambda\boldsymbol{x} = \boldsymbol{\Sigma}\boldsymbol{x} = \frac{\alpha}{d}\boldsymbol{x} + (1-\alpha)\boldsymbol{h}^\top\boldsymbol{x}\boldsymbol{h}. \tag{14}$$

Here, we consider the three cases as follows.

(i) the case where $\alpha = 1$. From (14), $\lambda_i = \frac{1}{d}$ for all $i = 1, \ldots, d$. Then, $\sum_{i=1}^d \lambda_i = 1$.

(ii) the case where $\alpha \in (0, 1)$ and $\|\boldsymbol{h}\| = 1$. From (14), $\boldsymbol{x}$ satisfies $\boldsymbol{x} = c\boldsymbol{h}$ for some constant $c \in \mathbb{R}$ or satisfies $\boldsymbol{h}^\top\boldsymbol{x} = 0$. In the case that $\boldsymbol{x} = c\boldsymbol{h}$, we have $\lambda = \frac{\alpha}{d} + (1-\alpha)$. In all other cases that $\boldsymbol{h}^\top\boldsymbol{x} = 0$, we have $\lambda = \frac{\alpha}{d}$. Then,

$$\sum_{i=1}^d \lambda_i = \frac{\alpha}{d} + (1-\alpha) + (d-1)\frac{\alpha}{d} = 1.$$

(iii) the case where $\alpha \in (0, 1)$ and $\boldsymbol{h} = \boldsymbol{0}$. From (14), we have $\lambda\boldsymbol{x} = \frac{\alpha}{d}\boldsymbol{x}$. Then,

$$\sum_{i=1}^d \lambda_i = \frac{\alpha}{d}d = \alpha \leq 1.$$

From the above discussion, (12) holds.

Regarding (13),

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^6] &\overset{(*)}{=} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[\left\|\sqrt{\frac{\alpha}{d}}\boldsymbol{w}+\sqrt{1-\alpha}\boldsymbol{h}v\right\|^6\right] \\
&\overset{(**)}{\leq} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[32\left\|\sqrt{\frac{\alpha}{d}}\boldsymbol{w}\right\|^6+32\left\|\sqrt{1-\alpha}\boldsymbol{h}v\right\|^6\right] \\
&\overset{(***)}{\leq} \frac{32\alpha^3}{d^3}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\|\boldsymbol{w}\|^6\right]+32(1-\alpha)^3\mathbb{E}_{v\sim\mathcal{N}(0,1)}[|v|^6] \\
&\overset{(****)}{\leq} \frac{32\alpha^3(d+6)^3}{d^3}+32(1-\alpha)^3 7^3 \\
&= \frac{32\alpha^3(d+6)^3}{d^3}+10976(1-\alpha)^3,
\end{aligned}$$

where (*) comes from Lemma D.4, (**) follows from Lemma D.6, (***) is due to the fact that $\|\boldsymbol{h}\|=1$ or $\|\boldsymbol{h}\|=0$, and (****) holds from (11) in Lemma D.1. □

Technical Lemma D.8. Suppose that Assumptions 4.2–4.4 hold. Then, $F_{\mu,\boldsymbol{\Sigma}}$ is H_F-smooth for any $\mu\in\mathbb{R}_{>0}$ and any positive-definite matrix $\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}$.

Proof.

$$\begin{aligned}
\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})-\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{y})\| &= \left\|\nabla\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[F(\boldsymbol{x}+\mu\boldsymbol{u})]-\nabla\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[F(\boldsymbol{y}+\mu\boldsymbol{u})]\right\| \\
&\overset{(*)}{=} \left\|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{x}+\mu\boldsymbol{u})]-\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{y}+\mu\boldsymbol{u})]\right\| \\
&\leq \left\|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{x}+\mu\boldsymbol{u})-\nabla F(\boldsymbol{y}+\mu\boldsymbol{u})]\right\| \\
&\leq \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[\|\nabla F(\boldsymbol{x}+\mu\boldsymbol{u})-\nabla F(\boldsymbol{y}+\mu\boldsymbol{u})\|] \\
&\overset{(**)}{\leq} \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})}[H_F\|\boldsymbol{x}-\boldsymbol{y}\|] \\
&= H_F\|\boldsymbol{x}-\boldsymbol{y}\|,
\end{aligned}$$

where (*) holds since Lemma D.2 enables to exchange the order of differentiation and integration. Moreover, (**) comes from Assumption 4.4. □

Technical Lemma D.9. Suppose that Assumptions 4.2 and 4.3 hold. Let

$$\boldsymbol{\Sigma}_1 := \frac{\alpha_1}{d}\boldsymbol{I}_d+(1-\alpha_1)\boldsymbol{h}_1\boldsymbol{h}_1^\top, \quad \boldsymbol{\Sigma}_2 := \frac{\alpha_2}{d}\boldsymbol{I}_d+(1-\alpha_2)\boldsymbol{h}_2\boldsymbol{h}_2^\top,$$

where $d\in\mathbb{N}$, $\alpha_1\in(0,1]$, $\alpha_2:=1-\gamma(1-\alpha_1)$ for $\gamma\in[0,1)$, $\boldsymbol{h}_1\in\mathbb{R}^d$ such that $\|\boldsymbol{h}_1\|=1$ or $\boldsymbol{h}_1=\boldsymbol{0}$, and $\boldsymbol{h}_2\in\mathbb{R}^d$ such that $\|\boldsymbol{h}_2\|=1$ or $\boldsymbol{h}_2=\boldsymbol{0}$. Then,

$$|F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})|\leq\sqrt{2}L_F|\mu_1-\mu_2|+L_F\mu_2|\sqrt{\alpha_1}-\sqrt{\alpha_2}|+L_F\mu_2(1+\sqrt{\gamma})\sqrt{1-\alpha_1}.$$

Proof. We have

$$\begin{aligned}
|F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})| &= |F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})+F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})| \\
&\leq |F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})|+|F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})|. \qquad (15)
\end{aligned}$$

Here,

$$\begin{aligned}
|F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x}) - F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})| &= \left|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}[F(\boldsymbol{x}+\mu_1\boldsymbol{u})] - \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}[F(\boldsymbol{x}+\mu_2\boldsymbol{u})]\right| \\
&= \left|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}[F(\boldsymbol{x}+\mu_1\boldsymbol{u}) - F(\boldsymbol{x}+\mu_2\boldsymbol{u})]\right| \\
&\leq \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}\left[|F(\boldsymbol{x}+\mu_1\boldsymbol{u}) - F(\boldsymbol{x}+\mu_2\boldsymbol{u})|\right] \\
&\overset{(*)}{\leq} \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}\left[L_F\left\|\mu_1\boldsymbol{u}-\mu_2\boldsymbol{u}\right\|\right] \\
&= L_F\left|\mu_1-\mu_2\right|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}[\|\boldsymbol{u}\|] \\
&\overset{(**)}{=} L_F\left|\mu_1-\mu_2\right|\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[\left\|\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\sqrt{1-\alpha_1}\boldsymbol{h}_1 v\right\|\right] \\
&\overset{(***)}{\leq} L_F\left|\mu_1-\mu_2\right|\left(\sqrt{\frac{\alpha_1}{d}}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{w}\|] + \sqrt{1-\alpha_1}\mathbb{E}_{v\sim\mathcal{N}(0,1)}[|v|]\right) \\
&\overset{(****)}{\leq} L_F\left|\mu_1-\mu_2\right|\left(\sqrt{\alpha_1}+\sqrt{1-\alpha_1}\right) \\
&\overset{(*****)}{\leq} \sqrt{2}L_F\left|\mu_1-\mu_2\right|, \qquad (16)
\end{aligned}$$

where (*) holds from Lemma D.2, (**) comes from Lemma D.4, (***) is due to the fact that $\|\boldsymbol{h}_1\| = 1$ or $\boldsymbol{h}_1 = \boldsymbol{0}$, (****) follows from Lemma D.1, and (*****) holds since $\sup_{a\in(0,1]}(\sqrt{a}+\sqrt{1-a}) = \sqrt{2}$.

Moreover,

$$\begin{aligned}
&|F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x}) - F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})| \\
&= \left|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_1)}[F(\boldsymbol{x}+\mu_2\boldsymbol{u})] - \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_2)}[F(\boldsymbol{x}+\mu_2\boldsymbol{u})]\right| \\
&\overset{(*)}{=} \left|\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v\right)\right]\right. \\
&\qquad \left.-\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right)\right]\right| \\
&\leq \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[\left|F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v\right) - F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right)\right|\right], \qquad (17)
\end{aligned}$$

where (*) comes from Lemma D.4. Here, from Lemma D.2,

$$\begin{aligned}
&\left|F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v\right) - F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right)\right| \\
&\leq L_F\left\|\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v - \mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}-\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right\| \\
&\leq L_F\left\|\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}-\mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}\right\| + L_F\left\|\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v-\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right\| \\
&= \frac{L_F\mu_2}{\sqrt{d}}\left|\sqrt{\alpha_1}-\sqrt{\alpha_2}\right|\|\boldsymbol{w}\| + L_F\mu_2\left\|\sqrt{1-\alpha_1}\boldsymbol{h}_1-\sqrt{1-\alpha_2}\boldsymbol{h}_2\right\||v|. \qquad (18)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\left\|\sqrt{1-\alpha_1}\boldsymbol{h}_1-\sqrt{1-\alpha_2}\boldsymbol{h}_2\right\| &\overset{(*)}{\leq} \sqrt{1-\alpha_1}+\sqrt{1-\alpha_2} \\
&\overset{(**)}{=} \sqrt{1-\alpha_1}+\sqrt{\gamma(1-\alpha_1)} \\
&= (1+\sqrt{\gamma})\sqrt{1-\alpha_1}, \qquad (19)
\end{aligned}$$

where (*) is due to the fact that $\|\boldsymbol{h}_1\| = 1$ or $\boldsymbol{h}_1 = \boldsymbol{0}$ and the fact that $\|\boldsymbol{h}_2\| = 1$ or $\boldsymbol{h}_2 = \boldsymbol{0}$. Moreover, (**) comes from the definition of α_2.

From (18) and (19),

$$\begin{aligned}&\left|F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_1}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_1}\boldsymbol{h}_1 v\right)-F\left(\boldsymbol{x}+\mu_2\sqrt{\frac{\alpha_2}{d}}\boldsymbol{w}+\mu_2\sqrt{1-\alpha_2}\boldsymbol{h}_2 v\right)\right|\\&\leq\frac{L_F\mu_2}{\sqrt{d}}\left|\sqrt{\alpha_1}-\sqrt{\alpha_2}\right|\|\boldsymbol{w}\|+L_F\mu_2(1+\sqrt{\gamma})\sqrt{1-\alpha_1}|v|.\end{aligned}\tag{20}$$

From (17) and (20), we have

$$\begin{aligned}&|F_{\mu_2,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})|\\&\leq\frac{L_F\mu_2}{\sqrt{d}}|\sqrt{\alpha_1}-\sqrt{\alpha_2}|\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{w}\|]+L_F\mu_2(1+\sqrt{\gamma})\sqrt{1-\alpha_1}\mathbb{E}_{v\sim\mathcal{N}(0,1)}[|v|]\\&\overset{(*)}{\leq}L_F\mu_2|\sqrt{\alpha_1}-\sqrt{\alpha_2}|+L_F\mu_2(1+\sqrt{\gamma})\sqrt{1-\alpha_1},\end{aligned}\tag{21}$$

where (*) follows from Lemma D.1.

From (15), (16), and (21), we have

$$|F_{\mu_1,\boldsymbol{\Sigma}_1}(\boldsymbol{x})-F_{\mu_2,\boldsymbol{\Sigma}_2}(\boldsymbol{x})|\leq\sqrt{2}L_F\,|\mu_1-\mu_2|+L_F\mu_2|\sqrt{\alpha_1}-\sqrt{\alpha_2}|+L_F\mu_2(1+\sqrt{\gamma})\sqrt{1-\alpha_1}.$$

□

Technical Lemma D.10. Suppose that F is Lipschitz continuous. For any $\boldsymbol{x}\in\mathbb{R}^d$, $\mu\in\mathbb{R}_{>0}$, and $\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}$, the following holds.

$$\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})=\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right]=\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right].$$

Proof. First, we show that, when $\int|F(\boldsymbol{y})|e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\mathrm{d}\boldsymbol{y}<\infty$,

$$\lim_{\boldsymbol{y}\to(\infty,\ldots,\infty)^\top}F(\boldsymbol{y})e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}=0,\quad\lim_{\boldsymbol{y}\to-(\infty,\ldots,\infty)^\top}F(\boldsymbol{y})e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}=0.$$

Assuming that the above does not hold, we derive a contradiction. Since

$$\lim_{\boldsymbol{y}\to(\infty,\ldots,\infty)^\top}F(\boldsymbol{y})e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\neq0,$$

there exist $c_1\in\mathbb{R}_{>0}$, $q\in\mathbb{R}_{>0}$, and $\{\boldsymbol{y}_n\}_{n=1}^\infty$ such that $\|\boldsymbol{y}_n-\boldsymbol{y}_{n+1}\|\geq q$ and $|F(\boldsymbol{y}_n)|e^{-\frac{1}{2\mu^2}(\boldsymbol{y}_n-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n-\boldsymbol{x})}\geq c_1$ for $n=1,\ldots,\infty$. Then, since F is Lipschitz continuous, there exist $c_2\in\mathbb{R}_{>0}$ and $\delta<\frac{q}{2}$ such that $|F(\boldsymbol{y})|e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\geq c_2$ for $\boldsymbol{y}_n-\delta\cdot\mathbf{1}\leq\boldsymbol{y}\leq\boldsymbol{y}_n+\delta\cdot\mathbf{1}$. Then, $\int_{\boldsymbol{y}_1-\delta\cdot\mathbf{1}}^\infty|F(\boldsymbol{y})|e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\mathrm{d}\boldsymbol{y}\geq\sum_{n=1}^\infty c_2(2\delta)^d=\infty$. It contradicts the fact that $\int|F(\boldsymbol{y})|e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\mathrm{d}\boldsymbol{y}<\infty$. Therefore,

$$\lim_{\boldsymbol{y}\to(\infty,\ldots,\infty)^\top}F(\boldsymbol{y})e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}=0.\tag{22}$$

Similarly,

$$\lim_{\boldsymbol{y}\to-(\infty,\ldots,\infty)^\top}F(\boldsymbol{y})e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}=0.\tag{23}$$

Let $\kappa := \sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}$ and $\boldsymbol{y} := \boldsymbol{x} + \mu\boldsymbol{u}$. Then, we have

$$\begin{aligned}
\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) &\overset{(*)}{=} \frac{1}{\kappa}\int \nabla_{\boldsymbol{x}} F(\boldsymbol{x}+\mu\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u} \\
&= \frac{1}{\mu^{d+2}\kappa}\int \nabla_{\boldsymbol{y}} F(\boldsymbol{y}) e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\mathrm{d}\boldsymbol{y} \\
&\overset{(**)}{=} \frac{1}{\mu^{d+2}\kappa}\left[F(\boldsymbol{y}) e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\right]_{\boldsymbol{y}\to-(\infty,\dots,\infty)^\top}^{\boldsymbol{y}\to(\infty,\dots,\infty)^\top} - \frac{1}{\mu^{d+2}\kappa}\int F(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\left(e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{y} \\
&\overset{(***)}{=} -\frac{1}{\mu^{d+2}\kappa}\int F(\boldsymbol{y}) e^{-\frac{1}{2\mu^2}(\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x})}(-\boldsymbol{\Sigma}^{-1})(\boldsymbol{y}-\boldsymbol{x})\mathrm{d}\boldsymbol{y} \\
&= \frac{1}{\mu^2\kappa}\boldsymbol{\Sigma}^{-1}\int F(\boldsymbol{x}+\mu\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{u}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}(\mu\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\
&= \frac{1}{\kappa}\boldsymbol{\Sigma}^{-1}\int \frac{F(\boldsymbol{x}+\mu\boldsymbol{u})}{\mu}\boldsymbol{u} e^{-\frac{1}{2}\boldsymbol{u}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u} \\
&= \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right], \qquad (24)
\end{aligned}$$

where (*) comes from the fact that F is Lipschitz continuous, (**) is due to integration by parts, and (***) follows from (22) and (23).

Moreover, since $\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right] = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}-\mu\boldsymbol{u})}{\mu}(-\boldsymbol{u})\right]$,

$$\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) = -\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}-\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right]. \qquad (25)$$

Then,

$$\begin{aligned}
\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) &= \frac{1}{2}\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) + \frac{1}{2}\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) \\
&\overset{(*)}{=} \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right] - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}-\mu\boldsymbol{u})}{\mu}\boldsymbol{u}\right] \\
&= \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right],
\end{aligned}$$

where (*) comes from (24) and (25). □

Technical Lemma D.11. Suppose that Assumption 4.4 holds. Let $\boldsymbol{\Sigma} := \alpha\frac{\boldsymbol{I}_d}{d} + (1-\alpha)\boldsymbol{h}\boldsymbol{h}^\top$ for $d\in\mathbb{N}$, $\boldsymbol{h}\in\mathbb{R}^d$ such that $\|\boldsymbol{h}\| = 1$ or $\boldsymbol{h} = \boldsymbol{0}$, and $\alpha\in(0,1]$. Then, for any $\boldsymbol{x}\in\mathbb{R}^d$ and $\mu\in\mathbb{R}_{>0}$,

$$\|\nabla F(\boldsymbol{x})\|^2 \le 2\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\|^2 + \frac{d^2\mu^2H_F^2}{\alpha^2}\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right).$$

Proof. First, we show that $\nabla F(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{x})^\top\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}]$. Let $\kappa := \int_{\mathbb{R}} e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u}$. Since $\kappa = \sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}$ and $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}^{-1}|^{-1}$,

$$\ln\left(\int_{\mathbb{R}} e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u}\right) = \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}^{-1}|).$$

Differentiating this identity in $\boldsymbol{\Sigma}^{-1}$, it follows from (Petersen et al., 2008, (57)) that

$$-\frac{1}{2\kappa}\int_{\mathbb{R}} \boldsymbol{u}\boldsymbol{u}^\top e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u} = -\frac{1}{2}\boldsymbol{\Sigma}^\top \overset{(*)}{=} -\frac{1}{2}\boldsymbol{\Sigma},$$

where (*) holds since $\boldsymbol{\Sigma}$ is a symmetric matrix. Thus, multiplying by $\nabla F(\boldsymbol{x})$ from the right yields

$$\frac{1}{\kappa}\int_{\mathbb{R}} \boldsymbol{u}\boldsymbol{u}^\top\nabla F(\boldsymbol{x}) e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u} = \boldsymbol{\Sigma}\nabla F(\boldsymbol{x}),$$

that is,

$$\nabla F(\boldsymbol{x}) = \frac{1}{\kappa}\int_{\mathbb{R}} \nabla F(\boldsymbol{x})^\top \boldsymbol{u}\boldsymbol{\Sigma}^{-1}\boldsymbol{u} e^{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u} = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{x})^\top \boldsymbol{u}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}].$$

Then,

$$\begin{aligned}
\|\nabla F(\boldsymbol{x})\|^2 &= \left\|\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\nabla F(\boldsymbol{x})^\top \boldsymbol{u}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}]\right\|^2 \\
&= \left\|\frac{1}{\mu}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\left(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}) - (F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}) - \mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u})\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right]\right\|^2 \\
&\le 2\left\|\frac{1}{\mu}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}))\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right]\right\|^2 \\
&\quad + 2\left\|\frac{1}{\mu}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}) - \mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u})\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right]\right\|^2 \\
&\overset{(*)}{=} 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + 2\left\|\frac{1}{\mu}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}) - \mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u})\boldsymbol{\Sigma}^{-1}\boldsymbol{u}]\right\|^2 \\
&\le 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + \frac{2}{\mu^2}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}) - \mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u})^2\|\boldsymbol{\Sigma}^{-1}\|^2\|\boldsymbol{u}\|^2] \\
&\overset{(**)}{\le} 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + \frac{2}{\mu^2}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{H_F^2}{4}\|\mu\boldsymbol{u}\|^4\|\boldsymbol{\Sigma}^{-1}\|^2\|\boldsymbol{u}\|^2\right] \\
&\overset{(***)}{=} 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + \frac{d^2\mu^2H_F^2}{2\alpha^2}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^6] \\
&\overset{(****)}{\le} 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + \frac{d^2\mu^2H_F^2}{2\alpha^2}\left(\frac{32\alpha^3(d+6)^3}{d^3} + 10976(1-\alpha)^3\right) \\
&= 2\left\|\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x})\right\|^2 + \frac{d^2\mu^2H_F^2}{\alpha^2}\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right),
\end{aligned}$$

where (*) comes from Lemma D.10 and the fact that $\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[F(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right] = F(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\boldsymbol{u}] = \boldsymbol{0}$, (**) is due to the fact that $F(\boldsymbol{x}+\mu\boldsymbol{u}) \le F(\boldsymbol{x}) + \nabla F(\boldsymbol{x})^\top(\mu\boldsymbol{u}) + \frac{1}{2}H_F\|\mu\boldsymbol{u}\|^2$ from Assumption 4.4, (***) follows from Lemma D.5, and (****) comes from Lemma D.7. □

Technical Lemma D.12. Suppose that Assumptions 4.2–4.4 holds. Let $\boldsymbol{\Sigma} := \alpha\frac{\boldsymbol{I}_d}{d} + (1-\alpha)\boldsymbol{h}\boldsymbol{h}^\top$ for $d \in \mathbb{N}$, $\boldsymbol{h} \in \mathbb{R}^d$ such that $\|\boldsymbol{h}\| = 1$ or $\boldsymbol{h} = \boldsymbol{0}$, and $\alpha \in (0,1]$. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\mu > 0$, the following holds.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right] \\
&\le \frac{8\alpha^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x})\|^2 + 8(1-\alpha)(25-23\alpha)\|\nabla F(\boldsymbol{x})\|^2 + \mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right).
\end{aligned}$$

Proof. Since F is H_F-smooth from Assumption 4.4,

$$\begin{aligned}
F(\boldsymbol{x}+\mu\boldsymbol{u}) &\le F(\boldsymbol{x}) + \nabla F(\boldsymbol{x})^\top(\mu\boldsymbol{u}) + \frac{1}{2}H_F\|\mu\boldsymbol{u}\|^2, \\
F(\boldsymbol{x}-\mu\boldsymbol{u}) &\ge F(\boldsymbol{x}) - \nabla F(\boldsymbol{x})^\top(\mu\boldsymbol{u}) - \frac{1}{2}H_F\|\mu\boldsymbol{u}\|^2.
\end{aligned}$$

Then, we have

$$\begin{aligned}
F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u}) &= [F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x})] + [F(\boldsymbol{x}) - F(\boldsymbol{x}-\mu\boldsymbol{u})] \\
&\le \left[\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} + \frac{\mu^2}{2}H_F\|\boldsymbol{u}\|^2\right] + \left[\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} + \frac{\mu^2}{2}H_F\|\boldsymbol{u}\|^2\right] \\
&= 2\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} + \mu^2H_F\|\boldsymbol{u}\|^2. \qquad (26)
\end{aligned}$$

Similarly, since

$$F(\boldsymbol{x}+\mu\boldsymbol{u}) \geq F(\boldsymbol{x}) + \nabla F(\boldsymbol{x})^\top(\mu\boldsymbol{u}) - \frac{1}{2}H_F\|\mu\boldsymbol{u}\|^2,$$
$$F(\boldsymbol{x}-\mu\boldsymbol{u}) \leq F(\boldsymbol{x}) - \nabla F(\boldsymbol{x})^\top(\mu\boldsymbol{u}) + \frac{1}{2}H_F\|\mu\boldsymbol{u}\|^2,$$

we have

$$F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u}) \geq 2\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} - \mu^2 H_F\|\boldsymbol{u}\|^2. \tag{27}$$

From (26) and (27),

$$\begin{aligned}(F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u}))^2 &\leq (2\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} + \mu^2 H_F\|\boldsymbol{u}\|^2)^2 + (2\mu\nabla F(\boldsymbol{x})^\top\boldsymbol{u} - \mu^2 H_F\|\boldsymbol{u}\|^2)^2\\ &= 8\mu^2(\nabla F(\boldsymbol{x})^\top\boldsymbol{u})^2 + 2\mu^4 H_F^2\|\boldsymbol{u}\|^4.\end{aligned} \tag{28}$$

Let $\rho := \nabla F(\boldsymbol{x})^\top\boldsymbol{h}$. Then,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u}) - F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right]\\
&= \frac{1}{4\mu^2}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[(F(x+\mu\boldsymbol{u}) - F(x-\mu\boldsymbol{u}))^2\|\boldsymbol{u}\|^2\right]\\
&\overset{(*)}{\leq} \frac{1}{4\mu^2}\left(\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[8\mu^2(\nabla F(\boldsymbol{x})^\top\boldsymbol{u})^2\|\boldsymbol{u}\|^2 + 2\mu^4 H_F^2\|\boldsymbol{u}\|^6]\right)\\
&\overset{(**)}{=} \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[2\left(\nabla F(x)^\top\left(\sqrt{\frac{\alpha}{d}}\boldsymbol{w} + \sqrt{1-\alpha}\boldsymbol{h}v\right)\right)^2\left\|\sqrt{\frac{\alpha}{d}}\boldsymbol{w} + \sqrt{1-\alpha}\boldsymbol{h}v\right\|^2\right] + \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{\mu^2 H_F^2}{2}\|\boldsymbol{u}\|^6\right]\\
&\overset{(***)}{\leq} 2\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[\left(\frac{2\alpha}{d}(\nabla F(x)^\top\boldsymbol{w})^2 + 2(1-\alpha)\rho^2v^2\right)\left(\frac{2\alpha}{d}\|\boldsymbol{w}\|^2 + 2(1-\alpha)v^2\right)\right]\\
&\quad + \frac{\mu^2 H_F^2}{2}\left(\frac{32\alpha^3(d+6)^3}{d^3} + 10976(1-\alpha)^3\right)\\
&= 2\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d),v\sim\mathcal{N}(0,1)}\left[\frac{4\alpha^2}{d^2}(\nabla F(x)^\top\boldsymbol{w})^2\|\boldsymbol{w}\|^2 + \frac{4\alpha(1-\alpha)}{d}(\nabla F(x)^\top\boldsymbol{w})^2v^2 + \frac{4(1-\alpha)\alpha}{d}\rho^2v^2\|\boldsymbol{w}\|^2\right.\\
&\qquad\qquad \left. + 4(1-\alpha)^2\rho^2v^4\right] + \mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right)\\
&\leq \frac{8\alpha^2}{d^2}\|\nabla F(x)\|^2\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{w}\|^4] + \frac{8\alpha(1-\alpha)\|\nabla F(\boldsymbol{x})\|^2}{d}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{w}\|^2]\mathbb{E}_{v\sim\mathcal{N}(0,1)}[v^2]\\
&\quad + \frac{8(1-\alpha)\alpha}{d}\rho^2\mathbb{E}_{v\sim\mathcal{N}(0,1)}[v^2]\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{w}\|^2]\\
&\quad + 8(1-\alpha)^2\rho^2\mathbb{E}_{v\sim\mathcal{N}(0,1)}[v^4] + \mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right)\\
&\overset{(****)}{\leq} \frac{8\alpha^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x})\|^2 + 8\alpha(1-\alpha)\|\nabla F(\boldsymbol{x})\|^2 + 8(1-\alpha)\alpha\|\nabla F(\boldsymbol{x})\|^2 + 8(1-\alpha)^2\|\nabla F(\boldsymbol{x})\|^2\cdot 25\\
&\quad + \mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right)\\
&= \frac{8\alpha^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x})\|^2 + 8(1-\alpha)(25-23\alpha)\|\nabla F(\boldsymbol{x})\|^2 + \mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3} + 5488(1-\alpha)^3\right),
\end{aligned}$$

where (*) comes from (28), (**) is due to Lemma D.4, (***) follows from Lemma D.7 and the fact that $\|\boldsymbol{h}\|^2 \leq 1$, and (****) comes from Lemma D.1 and the fact that $\rho = \nabla F(x)^\top\boldsymbol{h} \leq \|\nabla F(x)\|\|\boldsymbol{h}\| \leq \|\nabla F(x)\|$. □

D.2. Proofs of main lemmas and theorems in our paper

Here, we prove the main lemmas and theorems in our paper by using the technical lemmas prepared in Section D.1.

D.2.1. Proof of Lemma 3.2

Proof. From Lemma D.10,

$$\begin{aligned}\nabla F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) &= \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right]\\ &= \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\frac{\mathbb{E}_{\boldsymbol{\xi}^1\sim D(\boldsymbol{x}+\mu\boldsymbol{u})}[f(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{\xi}^1)]-\mathbb{E}_{\boldsymbol{\xi}^2\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}[f(\boldsymbol{x}-\mu\boldsymbol{u},\boldsymbol{\xi}^2)]}{2\mu}\boldsymbol{u}\right]\\ &= \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\mathbb{E}_{\boldsymbol{\xi}^1\sim D(\boldsymbol{x}+\mu\boldsymbol{u}),\boldsymbol{\xi}^2\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\left[\boldsymbol{g}(\boldsymbol{x},\mu,\boldsymbol{u},\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)\right]\right].\end{aligned}$$

□

D.2.2. Proof of Lemma 4.6

Proof. First, we show that $\|\boldsymbol{x}+\boldsymbol{y}+\boldsymbol{z}\|^2 \le 3\|\boldsymbol{x}\|^2+3\|\boldsymbol{y}\|^2+3\|\boldsymbol{z}\|^2$. The inequality of arithmetic and geometric means yields, for any $\boldsymbol{x}\in\mathbb{R}^d$, $\boldsymbol{y}\in\mathbb{R}^d$, and $\boldsymbol{z}\in\mathbb{R}^d$,

$$\|\boldsymbol{x}\|\|\boldsymbol{y}\| \le \frac{\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2}{2}, \quad \|\boldsymbol{y}\|\|\boldsymbol{z}\| \le \frac{\|\boldsymbol{y}\|^2+\|\boldsymbol{z}\|^2}{2}, \quad \|\boldsymbol{z}\|\|\boldsymbol{x}\| \le \frac{\|\boldsymbol{z}\|^2+\|\boldsymbol{x}\|^2}{2}.$$

Then,

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}+\boldsymbol{z}\|^2 &= \|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2+\|\boldsymbol{z}\|^2+2\boldsymbol{x}^\top\boldsymbol{y}+2\boldsymbol{y}^\top\boldsymbol{z}+2\boldsymbol{x}^\top\boldsymbol{z}\\ &\le \|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2+\|\boldsymbol{z}\|^2+2\|\boldsymbol{x}\|\|\boldsymbol{y}\|+2\|\boldsymbol{y}\|\|\boldsymbol{z}\|+2\|\boldsymbol{x}\|\|\boldsymbol{z}\|\\ &\le 3\|\boldsymbol{x}\|^2+3\|\boldsymbol{y}\|^2+3\|\boldsymbol{z}\|^2. \end{aligned} \tag{29}$$

Here,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m\sim D(\boldsymbol{x}+\mu\boldsymbol{u}),\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\left[\left\|\frac{1}{m}\sum_{j=1}^m \boldsymbol{g}(\boldsymbol{x},\mu,\boldsymbol{u},\boldsymbol{\xi}^{1,j},\boldsymbol{\xi}^{2,j})\right\|^2\right]\right]\\ &= \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m\sim D(\boldsymbol{x}+\mu\boldsymbol{u}),\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\left[\left\|\frac{\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{\xi}^{1,j})-\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}-\mu\boldsymbol{u},\boldsymbol{\xi}^{2,j})}{2\mu}\boldsymbol{u}\right\|^2\right]\right]\\ &= \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\Bigg[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m\sim D(\boldsymbol{x}+\mu\boldsymbol{u}),\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\Bigg[\Bigg\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}+\frac{\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{\xi}^{1,j})-F(\boldsymbol{x}+\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\\ &\qquad\qquad +\frac{-\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}-\mu\boldsymbol{u},\boldsymbol{\xi}^{2,j})+F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\Bigg\|^2\Bigg]\Bigg]\\ &\overset{(*)}{\le} \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\Bigg[3\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2+3\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m\sim D(\boldsymbol{x}+\mu\boldsymbol{u})}\left[\left\|\frac{\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}+\mu\boldsymbol{u},\boldsymbol{\xi}^{1,j})-F(\boldsymbol{x}+\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right]\\ &\qquad +3\mathbb{E}_{\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m\sim D(\boldsymbol{x}-\mu\boldsymbol{u})}\left[\left\|\frac{-\frac{1}{m}\sum_{j=1}^m f(\boldsymbol{x}-\mu\boldsymbol{u},\boldsymbol{\xi}^{2,j})+F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right]\Bigg]\\ &\overset{(**)}{\le} 3\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right]+\frac{3\sigma^2}{4\mu^2 m}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^2]+\frac{3\sigma^2}{4\mu^2 m}\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[\|\boldsymbol{u}\|^2]\\ &\overset{(***)}{\le} 3\mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\left[\left\|\frac{F(\boldsymbol{x}+\mu\boldsymbol{u})-F(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}\boldsymbol{u}\right\|^2\right]+\frac{3\sigma^2}{2\mu^2 m},\\ &\overset{(****)}{\le} \frac{24\alpha^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x})\|^2+24(1-\alpha)(25-23\alpha)\|\nabla F(\boldsymbol{x})\|^2+3\mu^2H_F^2\left(\frac{16\alpha^3(d+6)^3}{d^3}+5488(1-\alpha)^3\right)+\frac{3\sigma^2}{2\mu^2 m},\end{aligned}$$

where (*) follows from (29), (**) follows from Assumption 4.1 and Lemma D.3, (***) is due to (12) in Lemma D.7, and (****) comes from Lemma D.12.

□

D.2.3. PROOF OF THEOREM 4.7

Proof. From Lemma D.8, function $F_{\mu_k,\boldsymbol{\Sigma}_k}$ is H_F-smooth. Then,

$$\begin{aligned} F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1}) &\le F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) + \nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k) + \frac{H_F}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2 \\ &= F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) - \beta\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\boldsymbol{g}_k + \frac{H_F\beta^2}{2}\|\boldsymbol{g}_k\|^2. \end{aligned}$$

Here, let $\boldsymbol{\zeta}_k := (\boldsymbol{u}_k, \{\boldsymbol{\xi}_k^{1,j}\}_{j=1}^{m_k}, \{\boldsymbol{\xi}_k^{2,j}\}_{j=1}^{m_k}, \{\boldsymbol{\xi}_k^{3,j}\}_{j=1}^{n_k})$ and $\boldsymbol{\zeta}_{[0,k]} := (\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_k)$. Taking the expectation with respect to the random vectors $\boldsymbol{\zeta}_{[0,k]}$, we obtain

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1})] \\ &\le \mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}\left[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) - \beta\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\boldsymbol{g}_k + \frac{H_F\beta^2}{2}\|\boldsymbol{g}_k\|^2\right] \\ &= E_{\boldsymbol{\zeta}_{[0,k-1]}}\left[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) - \beta\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\mathbb{E}_{\boldsymbol{\zeta}_k}\left[\boldsymbol{g}_k \mid \boldsymbol{\zeta}_{[0,k-1]}\right] + \frac{H_F\beta^2}{2}\mathbb{E}_{\boldsymbol{\zeta}_k}\left[\|\boldsymbol{g}_k\|^2 \mid \boldsymbol{\zeta}_{[0,k-1]}\right]\right] \\ &\overset{(*)}{\le} E_{\boldsymbol{\zeta}_{[0,k-1]}}\Bigg[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) - \beta\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\boldsymbol{\Sigma}_k\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) + \frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x}_k)\|^2 \\ &\qquad + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\|\nabla F(\boldsymbol{x}_k)\|^2 + 3\mu_k^2H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2m_k}\Bigg]. \end{aligned} \tag{30}$$

where (*) comes from Lemmas 3.2 and 4.6. Here, when $\boldsymbol{s}_k \neq 0$, we have

$$\begin{aligned} \nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\boldsymbol{\Sigma}_k\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) &= \alpha_k d^{-1}\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2 + (1-\alpha_k)\|\boldsymbol{h}_k^\top\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2 \\ &\ge \alpha_k d^{-1}\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2. \end{aligned}$$

When $\boldsymbol{s}_k = 0$, we have

$$\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)^\top\boldsymbol{\Sigma}_k\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) = \alpha_k d^{-1}\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2.$$

Therefore, from (30) and the fact that $m_k \ge 1$,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1})] \\ &\le E_{\boldsymbol{\zeta}_{[0,k-1]}}\Bigg[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k) - \beta\alpha_k d^{-1}\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2 + \frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2\|\nabla F(\boldsymbol{x}_k)\|^2 \\ &\qquad + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\|\nabla F(\boldsymbol{x}_k)\|^2 + 3\mu_k^2H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2}\Bigg]. \end{aligned}$$

Rearranging the terms in the above inequality leads to the following.

$$
\begin{aligned}
&\beta\alpha_k d^{-1}\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2] \\
&\leq \mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] - \mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1})] \qquad (31)\\
&\quad + \left(\frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2 + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\quad + 3\mu_k^2 H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2} \\
&= \mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] + \mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[-F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1}) + F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1}) - F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})] \\
&\quad + \left(\frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2 + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\quad + 3\mu_k^2 H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2} \\
&\leq \mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] + \mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[|-F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_{k+1}) + F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})| - F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})] \\
&\quad + \left(\frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2 + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\quad + 3\mu_k^2 H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2} \\
&\overset{(*)}{\leq} \mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] - \mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})] + \sqrt{2}L_F(\mu_k-\mu_{k+1}) + L_F\mu_{k+1}(\sqrt{\alpha_{k+1}}-\sqrt{\alpha_k}) \\
&\quad + L_F\mu_{k+1}(1+\sqrt{\gamma})\sqrt{1-\alpha_k} + \left(\frac{12H_F\beta^2\alpha_k^2}{d^2}(d+4)^2 + 12H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\quad + 3\mu_k^2 H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{4\mu_k^2}, \qquad (32)
\end{aligned}
$$

where (*) holds from Lemma D.9 and the facts that $\mu_k \geq \mu_{k+1}$ and $\alpha_k \leq \alpha_{k+1}$ for all $k = 0, \ldots, T$.

Here, since it follows from Lemma D.11 that

$$
\|\nabla F(\boldsymbol{x}_k)\|^2 \leq 2\left\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\right\|^2 + \frac{d^2\mu_k^2 H_F^2}{\alpha_k^2}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right),
$$

we have

$$
\beta\alpha_k d^{-1}\|\nabla F(\boldsymbol{x}_k)\|^2 \leq 2\beta\alpha_k d^{-1}\left\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\right\|^2 + \frac{\beta d\mu_k^2 H_F^2}{\alpha_k}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right).
$$

Taking the expectation with respect to the random vectors $\boldsymbol{\zeta}_{[0,k-1]}$, we obtain

$$
\begin{aligned}
&\beta\alpha_k d^{-1}\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\leq 2\beta\alpha_k d^{-1}\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)\|^2] + \frac{\beta d\mu_k^2 H_F^2}{\alpha_k}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right) \\
&\overset{(*)}{\leq} 2\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] - 2\mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})] + 2\sqrt{2}L_F(\mu_k-\mu_{k+1}) + 2L_F\mu_{k+1}(\sqrt{\alpha_{k+1}}-\sqrt{\alpha_k}) \\
&\quad + 2L_F\mu_{k+1}(1+\sqrt{\gamma})\sqrt{1-\alpha_k} + \left(\frac{24H_F\beta^2\alpha_k^2(d+4)^2}{d^2} + 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2] \\
&\quad + 6\mu_k^2 H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) + \frac{3H_F\beta^2\sigma^2}{2\mu_k^2} + \frac{\beta d\mu_k^2 H_F^2}{\alpha_k}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right),
\end{aligned}
$$

where (*) follows from (32). Rearrange the terms in the above inequality, the following holds.

$$\begin{aligned}
&\left(\beta\alpha_k d^{-1} - \frac{24H_F\beta^2\alpha_k^2(d+4)^2}{d^2} - 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}\left[\|\nabla F(\boldsymbol{x}_k)\|^2\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[F_{\mu_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_k)] - 2\mathbb{E}_{\boldsymbol{\zeta}_{[0,k]}}[F_{\mu_{k+1},\boldsymbol{\Sigma}_{k+1}}(\boldsymbol{x}_{k+1})] + 2\sqrt{2}L_F(\mu_k-\mu_{k+1}) + 2L_F\mu_{k+1}(\sqrt{\alpha_{k+1}}-\sqrt{\alpha_k}) \\
&\quad + 2L_F\mu_{k+1}(1+\sqrt{\gamma})\sqrt{1-\alpha_k} + 6\mu_k^2H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) \\
&\quad + \frac{3H_F\beta^2\sigma^2}{2\mu_k^2} + \frac{\beta d\mu_k^2H_F^2}{\alpha_k}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right).
\end{aligned} \tag{33}$$

Here,

$$\begin{aligned}
\sum_{k=0}^{T} 2L_F\mu_{k+1}(1+\sqrt{\gamma})\sqrt{1-\alpha_k} &\overset{(*)}{\leq} \sum_{k=0}^{T} 2L_F\mu_0(1+\sqrt{\gamma})\sqrt{\gamma^k(1-\alpha_0)} \\
&= 2L_F\mu_0(1+\sqrt{\gamma})\sqrt{1-\alpha_0}\frac{1-\gamma^{\frac{T+1}{2}}}{1-\gamma^{\frac{1}{2}}} \\
&\leq \frac{2L_F\mu_0(1+\sqrt{\gamma})\sqrt{1-\alpha_0}}{1-\gamma^{\frac{1}{2}}},
\end{aligned} \tag{34}$$

where (*) comes from the updating rule of α_k and the fact that $\mu_k \leq \mu_0$ for $k = 1, \ldots, T$.

Let $F^* := \min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x})$. Then, summing up (33) for $0 \leq k \leq T$, we have

$$\begin{aligned}
&\sum_{k=0}^{T}\left(\beta\alpha_k d^{-1} - \frac{24H_F\beta^2\alpha_k^2(d+4)^2}{d^2} - 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k)\right)\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}\left[\|\nabla F(\boldsymbol{x}_k)\|^2\right] \\
&\overset{(*)}{\leq} 2F_{\mu_0,\boldsymbol{\Sigma}_0}(\boldsymbol{x}_0) - 2\mathbb{E}_{\boldsymbol{\zeta}_{[0,T]}}[F_{\mu_{T+1},\boldsymbol{\Sigma}_{T+1}}(\boldsymbol{x}_{T+1})] + 2\sqrt{2}(\mu_0-\mu_{T+1}) + 2L_F\mu_0(\sqrt{\alpha_{T+1}}-\sqrt{\alpha_0}) \\
&\quad + \frac{2L_F\mu_0(1+\sqrt{\gamma})\sqrt{1-\alpha_0}}{1-\gamma^{\frac{1}{2}}} + \sum_{k=0}^{T} 6\mu_0^2H_F^3\beta^2\left(\frac{8\alpha_k^3(d+6)^3}{d^3} + 2744(1-\alpha_k)^3\right) \\
&\quad + \sum_{k=0}^{T}\frac{3H_F\beta^2\sigma^2}{2\mu_k^2} + \sum_{k=0}^{T}\frac{\beta d\mu_0^2H_F^2}{\alpha_k}\left(\frac{16\alpha_k^3(d+6)^3}{d^3} + 5488(1-\alpha_k)^3\right) \\
&\overset{(**)}{\leq} 2F_{\mu_0,\boldsymbol{\Sigma}_0}(\boldsymbol{x}_0) - 2F^* + 2\sqrt{2}(\mu_0-\mu_{T+1}) + 2L_F\mu_0(\sqrt{\alpha_{T+1}}-\sqrt{\alpha_0}) + \frac{2L_F\mu_0(1+\sqrt{\gamma})\sqrt{1-\alpha_0}}{1-\gamma^{\frac{1}{2}}} \\
&\quad + 6(T+1)\mu_0^2H_F^3\beta^2\left(\frac{8(d+6)^3}{d^3} + 2744\right) + (T+1)\frac{3H_F\beta^2\sigma^2}{2\mu_{\min}^2} + (T+1)\frac{\beta d\mu_0^2H_F^2}{\alpha_0}\left(\frac{16(d+6)^3}{d^3} + 5488\right).
\end{aligned} \tag{35}$$

where (*) comes from (34) and $\mu_0 \geq \mu_k$ for all $k = 0, 1, \ldots, T$. The inequality (**) is due to that $\mu_{\min} \leq \mu_k$ for all $k = 0, 1, \ldots, T$, $0 < \alpha_0 \leq \alpha_k \leq 1$ for all $k = 0, 1, \ldots, T$, and $F_{\mu,\boldsymbol{\Sigma}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[F(\boldsymbol{x}+\mu\boldsymbol{u})] \geq \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}[F^*] = F^*$ for any $\boldsymbol{x} \in \mathbb{R}^d$, $\mu \geq 0$, and positive-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^d \times \mathbb{R}^d$.

Here, we have $\beta \leq \frac{\alpha_0 d}{48H_F((d+4)^2+25d^2)}$ and $0 \leq \alpha_0 \leq \alpha_k \leq 1$ from the assumptions on β and the updating rule of α_k. Then, for $k = 0, 1, \ldots, T$,

$$\begin{aligned}
&\beta\alpha_k d^{-1} - 24H_F\beta^2\alpha_k^2(d+4)^2d^{-2} - 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k) \\
&\geq \beta\alpha_0 d^{-1} - 24H_F\beta^2(d+4)^2d^{-2} - 24H_F\beta^2\cdot 25 \\
&= \beta\alpha_0 d^{-1} - 24H_F\beta^2d^{-2}((d+4)^2+25d^2) \\
&\geq \beta\alpha_0 d^{-1} - 24H_F\beta d^{-2}((d+4)^2+25d^2)\frac{\alpha_0 d}{48H_F((d+4)^2+25d^2)} \\
&= \frac{\beta\alpha_0}{2d}.
\end{aligned} \tag{36}$$

Therefore, from the definition of $\boldsymbol{x}_R$,

$$
\begin{aligned}
&\mathbb{E}_{R,\boldsymbol{\zeta}_{[0,T]}}[\|\nabla F(\boldsymbol{x}_R)\|^2] \\
&= \frac{\sum_{k=0}^{T}(\beta\alpha_k d^{-1} - 24H_F\beta^2\alpha_k^2(d+4)^2d^{-2} - 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k))\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2]}{\sum_{\ell=0}^{T}(\beta\alpha_\ell d^{-1} - 24H_F\beta^2\alpha_\ell^2(d+4)^2d^{-2} - 24H_F\beta^2(1-\alpha_\ell)(25-23\alpha_\ell)))} \\
&\overset{(*)}{\leq} \frac{2d(\sum_{k=0}^{T}(\beta\alpha_k d^{-1} - 24H_F\beta^2\alpha_k^2(d+4)^2d^{-2} - 24H_F\beta^2(1-\alpha_k)(25-23\alpha_k))\mathbb{E}_{\boldsymbol{\zeta}_{[0,k-1]}}[\|\nabla F(\boldsymbol{x}_k)\|^2])}{(T+1)\beta\alpha_0} \\
&\overset{(**)}{\leq} \frac{4d\left(F_{\mu_0,\boldsymbol{\Sigma}_0}(\boldsymbol{x}_0) - F^* + \sqrt{2}L_F(\mu_0-\mu_{T+1}) + (\sqrt{\alpha_{T+1}}-\sqrt{\alpha_0})L_F\mu_0 + L_F\mu_0(1+\sqrt{\gamma})\sqrt{1-\alpha_0}\left(1-\gamma^{\frac{1}{2}}\right)^{-1}\right)}{(T+1)\beta\alpha_0} \\
&\quad + \frac{12d\mu_0^2H_F^3\beta}{\alpha_0}\left(\frac{8(d+6)^3}{d^3}+2744\right) + \frac{3dH_F\beta\sigma^2}{\mu_{\min}^2\alpha_0} + \frac{2d^2\mu_0^2H_F^2}{\alpha_0^2}\left(\frac{16(d+6)^3}{d^3}+5488\right) \\
&= O((1+\mu_0)T^{-1}\beta^{-1}d) + O(\mu_0^2\beta d) + O(\beta\mu_{\min}^{-2}d\sigma^2) + O(d^2\mu_0^2),
\end{aligned}
$$

where (*) comes from (36) and (**) follows from (35). Here, $\mu_0 = \Theta(\epsilon d^{-1})$, $\mu_{\min} = \Theta(\epsilon d^{-1})$ from the assumption. Moreover, since $\beta = \min\left(\frac{\alpha_0 d}{48H_F((d+4)^2+25d^2)}, T^{-\frac{2}{3}}d^{-\frac{1}{3}}\right)$, we have

$$
\begin{aligned}
&\beta = O(T^{-\frac{2}{3}}d^{-\frac{1}{3}}), \text{ and} \\
&\beta^{-1} \leq \frac{48H_F((d+4)^2+25d^2)}{\alpha_0 d} + T^{\frac{2}{3}}d^{\frac{1}{3}} = O(d + T^{\frac{2}{3}}d^{\frac{1}{3}}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{R,\boldsymbol{\zeta}_{[0,T]}}[\|\nabla F(\boldsymbol{x}_R)\|^2] \\
&= O\left((1+\epsilon d^{-1})(T^{-1}d^2 + T^{-\frac{1}{3}}d^{\frac{4}{3}})\right) + O(\epsilon^2T^{-\frac{2}{3}}d^{-\frac{4}{3}}) + O(T^{-\frac{2}{3}}d^{\frac{8}{3}}\epsilon^{-2}\sigma^2) + O(\epsilon^2).
\end{aligned}
$$

By setting $T = \Theta(\sigma^3\epsilon^{-6}d^4)$, we obtain

$$
\mathbb{E}_{R,\boldsymbol{\zeta}_{[0,T]}}[\|\nabla F(\boldsymbol{x}_R)\|^2] \leq \epsilon^2.
$$

Here, the sample complexity is $O(\sigma^3\epsilon^{-6}d^4)$ since $\sum_{k=0}^{T}(2m_k+n_k) = \sum_{k=0}^{T}O(1) = O(T) = O(\sigma^3\epsilon^{-6}d^4)$, where $2m_k+n_k$ is the number of samples at the k-th iteration.

□

D.2.4. PROOF OF THEOREM 4.8

Proof. Since the proof of Theorem 4.7 does not depend on the value of $\boldsymbol{s}$ in Algorithm 1, the same iteration complexity holds for Algorithm 2. That is, by setting $T = \Theta(\sigma^3\epsilon^{-6}d^4)$, we obtain $\mathbb{E}_{R,\boldsymbol{\zeta}_{[0,T]}}[\|\nabla F(\boldsymbol{x}_R)\|^2] \leq \epsilon^2$. Then, the sample complexity of Algorithm 2 is $O(\sigma^3\epsilon^{-6}d^4)$ since $\sum_{k=0}^{T}2m_k = \sum_{k=0}^{T}O(1) = O(T) = O(\sigma^3\epsilon^{-6}d^4)$, where $2m_k$ is the number of samples at the k-th iteration.

□